

LUIS PRIETO VALIENTE
INMACULADA HERRANZ TEJEDOR

¿QUÉ SIGNIFICA “ESTADÍSTICAMENTE SIGNIFICATIVO”?

LA FALACIA DEL CRITERIO DEL 5%
EN LA INVESTIGACIÓN CIENTÍFICA

“Adiós “P-menor-del-0.05”, equívoco y traicionero
compañero de viaje. Tus efectos colaterales y toxicidad
intracerebral son demasiado grandes para compensar
cualquier beneficio que pudieras aportar”

A. R. Feinstein (1985) “Clinical Epidemiology: the architecture
of clinical research”, Philadelphia, WB saunders.

Con cuestionarios de autoevaluación para que el lector
pueda valorar su nivel de conocimiento



Subido por:



Interfase IQ

Libros de Ingeniería Química y más



<https://www.facebook.com/pages/Interfase-IQ/146073555478947?ref=bookmarks>

**Si te gusta este libro y tienes la posibilidad,
cómpralo para apoyar al autor.**

Este libro sale a la luz con la vocación de llegar al mayor número de lectores posible. Por ello el que una empresa de la dimensión de Schering-Plough decida colaborar a su difusión entre los profesionales de las Ciencias de la Salud es una ayuda decisiva. Al establecer un puente entre autores y lectores, Schering-Plough insiste en su ya tradicional línea de apoyo a la formación continuada de quienes trabajan día a día en las ciencias de la vida. Esperamos que esta obra, pensada para ellos, les ayude eficazmente a entender mejor las conclusiones que encuentran en las revistas científicas y a elaborar las de sus propios trabajos.

LUIS PRIETO e INMACULADA HERRANZ
Madrid, a 20 de enero del 2005

Luis Prieto Valiente
Inmaculada Herranz Tejedor

Profesores de Metodología de la Investigación en la Universidad Complutense
de Madrid

¿QUÉ SIGNIFICA «ESTADÍSTICAMENTE SIGNIFICATIVO»?

**La falacia del criterio del 5%
en la investigación científica**

© Luis Prieto Valiente, Inmaculada Herranz Tejedor
2005 (Libro en papel)
2015 (Libro electrónico)

Reservados todos los derechos.

“No está permitida la reproducción total o parcial de este libro, ni su tratamiento informático, ni la transmisión de ninguna forma o por cualquier medio, ya sea electrónico, mecánico, por fotocopia, por registro u otros métodos sin el permiso previo y por escrito de los titulares del Copyright”

Ediciones Díaz de Santos, S.A.
Albasanz, 2
28037 MADRID

ediciones@editdiazdesantos.com
www.editdiazdesantos.com

ISBN: 978-84-9969-950-9 (Libro electrónico)
ISBN: 978-84-7978-666-3 (Libro en papel)

Agradecimientos

En la redacción de este libro ha sido decisiva la ayuda de nuestros alumnos, colegas y maestros, muy en especial:

Jaime y Alfonso

Blanca, Ana Belén y Lucía.

Marga y Mario.

José Luis y Araceli,
los primeros y mejores maestros.

Los profesores Inmaculada Marrero, José María Martín, Rafael Alonso y Rafael Romero, que accedieron a escribir los prólogos avalando nuestras hipótesis.

Doctor Juan Carlos Diezma, corrector riguroso en estilo y contenido.

Doctor Manuel Martínez-Sellés y Doctor David Prieto, a los que ayer enseñamos y de los que hoy aprendemos.

Doctor Francisco R. Salvanés, de quien aprendemos permanentemente.

Profesora M.^a Luisa Martínez-Frías, con quien compartimos el empeño en explicar con claridad la Inferencia Estadística a todos nuestros colegas interesados en investigación. Su constante ayuda ha sido determinante.

Sergio Pequeño Ciaurritz, maestro y amigo.

De todos ellos esperamos que continúen ayudándonos a alcanzar el objetivo que este libro persigue.

Nuestro agradecimiento a D. José Manuel Díaz por facilitarnos en todo momento la labor de edición.

Índice

Presentación	XI
Prólogo	XIII
Capítulo 1. ¿Por qué este informe?	1
Capítulo 2. El testimonio de los expertos	7
Capítulo 3. Los errores más graves y frecuentes	13
Capítulo 4. La inferencia en la vida común	19
Capítulo 5. La Inferencia Estadística en investigación médica ..	29
Capítulo 6. Interpretación del valor P de los tests de significación	41
Capítulo 7. Probabilidad de un valor particular <i>versus</i> probabilidad de cola	53
Capítulo 8. Más ejemplos de interpretación del valor P del test ..	61
Capítulo 9. Tests de significación comparando dos medias y dos proporciones	67
Capítulo 10. No afirmar la hipótesis nula	77
Capítulo 11. La falsa frontera del 5%	85
Capítulo 12. El origen del malentendido: pensar <i>versus</i> decidir ..	95
Capítulo 13. Test de significación <i>versus</i> test de hipótesis	101
Capítulo 14. Lo que no es el valor P del test	117
Capítulo 15. El enigma del tamaño de la muestra	125
Conclusiones	137

Apéndices

1. Encuestas de autoevaluación previas	141
2. Encuestas de autoevaluación específicas	147
3. Soluciones a las encuestas de autoevaluación	167
4. Comentarios del Prof. Rafael Romero Villafranca	169

Presentación

Despierte el alma dormida,
avive el seso y recuerde...

que los tests estadísticos fueron creados para ayudar al investigador a elaborar conclusiones más razonables, lo cual no se consigue aplicando mecánicamente recetas como « $P < 0,05$, resultado estadísticamente significativo», cuyo significado real no entiende en muchos casos ni quien las escribe ni quien las lee.

Es lamentable ver a investigadores, tanto jóvenes como veteranos muy cualificados en su campo, diciendo arbitrariedades al interpretar los tests estadísticos de sus trabajos. Y lo es más aún verles refugiarse con fe ciega en la llamada «regla del 5%», que atribuye a esa cantidad propiedades mágicas.

Si toda regla rígida es inapropiada en el quehacer científico, esta es una de las más extendidas y nefastas. No hay nada limítrofe en el valor $P = 0,05$, ni en el $0,01$ ni en ninguna otra cantidad concreta. No hay un valor de P frontera que separe los resultados «válidos» de los «no válidos».

Es hora de decir con Moyé (2000): «Debemos reconocer que $P < 0,05$ se ha convertido en el estándar, pero ha llegado el momento de decir que eso es malo. Decidir mecánicamente de acuerdo al valor de P es una miope renuncia a razonar. Francamente, si usted desea que los valores P sustituyan a su capacidad pensante, probablemente ha llegado el momento de que usted deje esta actividad. Ya hemos superado el “ $P < 0,05$ ”. Hemos estado rindiéndole pleitesía demasiado tiempo. Ha llegado el momento de rebelarse contra la tiranía».

Y es hora de explicar a todos los investigadores que aunque los tests estadísticos son una cuestión matemática y a los estadísticos se debe su desarrollo teórico y su aplicación práctica, pueden explicarse, entenderse y usarse correctamente sin recurrir a noción matemática alguna. Todo lo

que se necesita es aplicar correctamente el sentido común, la lógica básica propia de todos los humanos.

Acerca de las autoevaluaciones (Apéndices 1 a 3)

Tras el primer capítulo, que justifica la salida del libro al mercado, se invita al lector a hacer las pruebas de autoevaluación del Apéndice 1.

Tienen por objeto ayudarle a ver si conoce realmente los conceptos que la obra aborda, o bien necesita clarificarlos. Nuestra sugerencia es que intente responder a las cuestiones que se plantean y elabore su propia puntuación en una escala de 1 a 10 siguiendo el sencillo procedimiento indicado.

En los capítulos cuatro y siguientes se desarrolla cada uno de los puntos que el investigador necesita conocer para elaborar correctamente las conclusiones de sus trabajos y entender las de sus colegas. Para muchos de esos capítulos también se han confeccionado autoevaluaciones específicas que van incluidas en el Apéndice 2 y ayudarán al lector a comprobar que ha asimilado los conceptos expuestos en ese capítulo.

Prólogo

Una posible clasificación de los investigadores en Ciencias de la Salud contemplaría cuatro grandes familias: a) clínicos, b) básicos, c) epidemiólogos y d) especialistas en salud pública y gestión de recursos sanitarios.

Con todos ellos colaboran los estadísticos que trabajan en contacto directo con la investigación.

Hemos pedido a profesionales muy cualificados en cada uno de esos campos un pequeño prólogo, para que cada lector pueda sentirse representado en uno de ellos. El primer prólogo, el Apéndice final y algunas citas del segundo capítulo, se lo debemos a un profesor de estadística, con larga experiencia en investigación y aplicaciones.

ESTADÍSTICA

Aunque la lógica de las ideas aquí desarrolladas las defiende por sí mismas, sin necesidad de valedores, no he querido negarme a la invitación que los autores me hacen para presentar su obra, bien entendido que mis opiniones no tienen más valor que el ser emitidas tras 40 años dedicados a la docencia, consultoría e investigación en Estadística aplicada.

Las ideas que se exponen a lo largo de esta obra son de gran importancia para cualquier persona que utilice técnicas estadísticas en su trabajo y por ello doy la bienvenida a este libro, y agradezco a los autores su esfuerzo para aclarar la verdadera naturaleza de los métodos estadísticos, contribuyendo con ello a poner nuestra ciencia al servicio de la sociedad.

RAFAEL ROMERO VILLAFRANCA

Catedrático de Estadística de la Universidad Politécnica de Valencia

INVESTIGACIÓN BÁSICA

Debo a Luis Prieto, con quien me inicié en ciertas aventuras intelectuales allá por los orígenes de casi todo, un razonable conocimiento acerca de los misterios de la P y la banalidad de la división en torno al 5%. En Inmaculada Herranz admiro la claridad y energía con que explica estos temas, haciéndolos accesibles a alumnos de muy diversas profesiones y edades.

Porque la Inferencia Estadística es paso obligado —y muy útil— para los investigadores de todas las ciencias, es realmente preocupante que muchos de mis colegas sigan sin entender sus conceptos básicos. Por eso, como editor del boletín de la Sociedad Española de Ciencias Fisiológicas, pedí a Luis e Inmaculada que escribiesen un artículo periódico, divulgando entre los fisiólogos las ideas básicas de los tests de significación. Es para mi una satisfacción ver que los artículos son elogiados y seguidos con sumo interés.

Cuando en 2002 organicé el primer congreso conjunto entre las sociedades española e inglesa de Fisiología, pensé que la correcta interpretación de los tests estadísticos constituía para nosotros un reto pendiente que requería tratamiento monográfico. Nuestros colegas británicos, con extensa tradición en este campo, estuvieron plenamente de acuerdo y corroboraron mi invitación a Luis para que expusiera este tema a profesionales sin formación matemática. La sesión resultó un éxito y generó ulteriores actividades en esa línea.

Luis e Inmaculada llevan décadas trabajando duro para que los investigadores biomédicos accedan a unos conocimientos que ellos explican con extrema claridad. Muchas generaciones de estudiantes e investigadores han aprovechado sus enseñanzas *in vivo*. Ahora este libro pone esas enseñanzas al alcance de todos, y lo hace de manera sencilla y sumamente asequible, mostrando que el razonamiento utilizado en los tests estadísticos es exactamente igual que el usado por todo contribuyente en su casa y en la calle.

Enhorabuena a los autores por haberlo conseguido y a los futuros lectores por la oportunidad de disfrutar y aprender que aquí tienen.

RAFAEL ALONSO SOLÍS

*Catedrático de Fisiología de la Universidad de La Laguna
Presidente Electo de la Sociedad Española de Ciencias Fisiológicas*

SALUD PÚBLICA Y ADMINISTRACIÓN DE RECURSOS SANITARIOS

Durante el período en el que tuve el privilegio de ser Director de la Escuela Nacional de Sanidad (1995-2000), conocí a Luis Prieto e Inmaculada Herranz, que se distinguían por su capacidad para hacer llegar a todos los profesionales, de forma amena y al mismo tiempo rigurosa, las ideas fundamentales del análisis estadístico.

Haciendo encuestas sistemáticas a los alumnos para ver si tenían claros los fundamentos de los tests estadísticos, constataron una y otra vez que muchos investigadores tenían ideas ciertamente equívocas. Y como no se resignan a dejar sumidos en la confusión y cautivos de recetas absurdas a estudiantes y profesionales, imparten continuamente conferencias sobre el tema. Saben que se puede entender la esencia de los tests estadísticos sin tener conocimientos matemáticos especializados y están empeñados en que esa realidad llegue a todos los científicos relacionadas con la salud y las ciencias psicológicas y sociales.

Comprender bien el sentido y fundamento de los tests estadísticos es esencial en salud pública y gestión de servicios sanitarios. Durante mi experiencia como Director General de Salud Pública del Ministerio de Sanidad (2002-2004) pude comprobar que el comprender de forma fehaciente qué significa que la exposición a los residuos del Prestige conlleva (o no) un «significativo» aumento del riesgo... o que la ola de calor ha hecho aumentar la mortalidad de forma «significativa»... tiene consecuencias prácticas evidentes si queremos proceder con rigor.

Estoy convencido de la trascendencia del tema y de la capacidad de los autores para explicarlo. Y confío realmente en que este libro contribuya a que la Inferencia Estadística aplicada a la investigación en ciencias de la vida deje de ser la asignatura pendiente de muchos estudiantes y estudiosos, y pase a ser materia de la que se disfrute, permitiendo que sea aprobada con brillantez y utilizada con inteligencia.

JOSÉ MARÍA MARTÍN MORENO

*Catedrático de Medicina Preventiva y Salud Pública
de la Universidad de Valencia*

Adviser, Health Intelligence. World Health Organization

EPIDEMIOLOGÍA

A principios de los 70, empecé a organizar el «Estudio Colaborativo Español de Malformaciones Congénitas» (ECEMC), diseño Caso-Control que inspecciona cada año más del 26% de los nacimientos nacionales y actualmente contiene más de 66.000 registros, con 300 variables para cada uno. En esos años era difícil hacer entender la necesidad del análisis estadístico en el ámbito de la medicina y era casi imposible encontrar un profesional que supiera aplicarlo a datos médicos. Por ello, fue para mí una gran suerte conocer a Luis Prieto, que había completado su formación con el Prof. P. Armitage, en el Dpto. de Biomatemáticas de la Universidad de Oxford, y había puesto en marcha el primer servicio de Bioestadística hospitalaria de nuestro país, en la Universidad de la Laguna.

Lo que más me impactó de Luis fue su capacidad para transmitir los conceptos estadísticos a todo tipo de alumnos y su permanente vocación de estudio para resolver los nuevos retos que yo constantemente le ponía. Ello hizo posible la publicación de varios trabajos con aportaciones significativas en revistas científicas americanas. Por otra parte, sus artículos de divulgación estadística para pediatras, dismorfólogos y genetistas son leídos con provecho por muchos de nuestros colegas.

Decía Sir Hunphry Davy que su mejor descubrimiento había sido Faraday. Salvando las distancias, yo diría que mi mejor aportación a la investigación epidemiológica ha sido Luis Prieto y que una de sus mejores aportaciones a la docencia de esta materia, ha sido Inmaculada Herranz. Inmaculada empezó su andadura en el mundo de la Bioestadística como becaria del ECEMC. Tiene una notable capacidad para transmitir claridad en sus explicaciones, siendo una de las pocas matemáticas capaces de explicar los conceptos de Estadística en un lenguaje que entienden todos los investigadores. Hace fácil lo difícil, tanto cuando explica fundamentos lógicos del análisis estadístico como cuando enseña a usarlo con recursos informáticos.

Este libro consigue hacernos ver que no compete al investigador tomar decisiones (sino matizar opiniones), que la mayoría de sus trabajos no le proporcionarán evidencias definitivas y que autor y lector deben valorar adecuadamente las evidencias parciales que cada trabajo aporta. Supondrá una gran ayuda para todos los que necesitan clarificar las bases

del razonamiento que rige esta materia, incluyendo algunos editores y revisores de revistas biomédicas de prestigio (nacionales e internacionales), que evalúan trabajos epidemiológicos haciendo una incorrecta interpretación de la Inferencia Estadística. Tanto es así que en la literatura científica reciente hay trabajos críticos sobre esta situación. Esto hace que el presente libro sea no sólo útil, sino que esté de total actualidad.

MARÍA LUISA MARTÍNEZ-FRÍAS

Profesora de la Facultad de Medicina de la Universidad

Complutense de Madrid

Directora del Centro de Investigación sobre Anomalías Congénitas

(CIAC) Instituto de Salud Carlos III

MEDICINA ASISTENCIAL

Conocí a Luis Prieto cuando iniciaba su carrera como profesor en la Facultad de Medicina de La Laguna. Su pasión por la lógica, su afán por hacernos pensar y no dejarnos impresionar por los números y los supuestos dogmas científicos han influido decisivamente en aquellos de sus alumnos que nos iniciábamos tímidamente en el mundo de la investigación. Y supe de Inmaculada Herranz, porque de su excepcional capacidad docente dan testimonio miles de estudiantes pre y post graduados.

Para los que nos dedicamos a la clínica, es un regalo inapreciable que se nos ayude a valorar sensatamente hasta qué punto los datos de nuestra investigación son o no compatibles con las hipótesis planteadas. Es un ejercicio que parece sencillo, pero no es habitual en los trabajos que leemos en revistas profesionales. Liberarse del corsé del valor de P y disfrutar de un trabajo bien planteado, honestamente realizado y con un análisis estadístico correcto, es un placer que recomiendo a todos los compañeros de profesión que quieran añadir un aliciente extra a la tarea asistencial del día a día.

También nuestros pacientes de cada mañana se beneficiarán de esta tarea, necesaria para ir un poco más allá en el conocimiento de nuestro trabajo. Y nosotros nos beneficiaremos de este libro que sus autores sacan a la luz porque se niegan a permanecer indiferentes viendo que

miles de investigadores tenemos problemas al interpretar el valor P de los tests, nos sentimos inseguros al respecto y acabamos refugiándonos en recetas grotescas.

Luis e Inmaculada saben que todos podríamos tener ideas claras y sentirnos cómodos en esta cuestión. Leyendo este libro se comprende que las cosas son más sencillas de lo que parecen si aprendemos a verlas con la mirada ingenua y la mente despierta.

INMACULADA MARRERO DOMÍNGUEZ
Hospital Universitario Insular de Gran Canaria

¿Por qué este informe?

Este informe se hacía necesario por tres motivos fundamentales.

1. La mayoría de los investigadores tienen serios problemas al interpretar el valor P de los tests estadísticos

Para la mayoría de los investigadores de las Ciencias de la Salud (y de todas las ciencias experimentales) entender el valor P de los tests estadísticos es la eterna asignatura pendiente. Asisten a cursillos, compran libros, leen artículos... pero no acaban de tener clara la situación y siguen cometiendo notables imprecisiones al elaborar las conclusiones de sus trabajos a partir de ese valor.

Todos ellos conviven constantemente con el problema de la Inferencia: ¿Hasta qué punto el hallazgo encontrado en la muestra analizada es una verdad general, válida para toda la población? ¿Qué resultados son mera anécdota particular de la muestra y cuáles son válidos más allá de ella?. Puesto que el investigador observa *muestras* que son solo una pequeña parte de la *población* que intenta conocer, esa pregunta es constante en su actividad científica, cualquiera que sea su campo.

Hasta finales del XIX no contaba con el apoyo del cálculo de probabilidades para ayudarle en este problema, pero el siglo XX trae consigo un gran regalo para los profesionales de las ciencias experimentales. Pearson, Student y Fisher desarrollan las bases de la Inferencia Estadística y ponen a disposición de los investigadores dos herramientas fundamentales: los tests estadísticos y los intervalos de confianza.

Sin embargo, el beneficio que podían proporcionar esos nuevos recursos se vio sensiblemente mermado por el escaso uso de los intervalos de confianza y el uso excesivo y muchas veces incorrecto de los tests estadísticos.

La mayoría de los investigadores tienen graves dificultades para entender lo que indica el valor P de los tests y depositan toda su confianza en una superstición tan extendida como absurda: la «regla del 5%». Creen que $P < 0.05$ garantiza que el tipo de resultado encontrado en la muestra es una verdad universal, de modo que:

«Resultado significativo» ($P < 0.05$) → lo encontrado en la muestra es extrapolable a la población.

«Resultado no significativo» ($P > 0.05$) → lo encontrado en la muestra no es extrapolable a la población.

Más del 90% de los investigadores asumen este criterio con inquebrantable fe y al analizar los resultados de sus investigaciones sus expectativas se centran obsesivamente en que «la P sea menor de 0.05». No saben por qué, pero saben que si es $P < 0.05$ su trabajo será más publicado, leído, valorado y citado. Se aferran a esta pintoresca superstición, desoyendo, además del sentido común, las muchas voces que continuamente les advierten. Los más inexpertos pueden, pasar de la desolación a la euforia si el valor P de su estudio pasa, por ejemplo, de 0.051 a 0.049.

El problema es más grave si estos profesionales no son conscientes de sus serias lagunas en este campo.

2. Evitar esos errores no requiere conocer el fundamento matemático de los tests estadísticos

La errónea interpretación de los tests estadísticos no refleja una especial incapacidad de los investigadores biológicos, sino la deficiente enseñanza que la mayoría de ellos recibieron en este campo. Siendo la Inferencia Estadística una disciplina matemática, suele ser explicada a los profesionales de otras ciencias usando un lenguaje matemático que la hace ininteligible para la mayoría, aunque todos ellos tienen sobrada capacidad para entender sin ambigüedad lo que el valor P indica, no sentirse perdidos en ese tema y no tener que recurrir a burdas simplificaciones.

Entender lo que el valor P de los tests indica es un tema meramente lógico al alcance de todo profesional. Se trata de un razonamiento que todos usamos en la vida diaria y se puede explicar apoyándose en ejemplos prácticos, apelando únicamente a la intuición y la lógica común, sin usar herramienta matemática alguna, como se aprende a interpretar una imagen radiológica sin ser experto en radiaciones, a conducir un coche sin necesidad de estudiar termodinámica, o a usar un ordenador sin ser experto en electrónica.

De hecho, algunos de los más cualificados estadísticos han escrito excelentes tratados exponiendo lo fundamental de estos conocimientos, y de su uso práctico, sin utilizar más aparato matemático que las reglas de la aritmética elemental. Los textos de Yule y Kendall (1953), Fisher (1925, 1935 y 1956), Snedecord y Cochran (1950), Cochran y Cox (1957), Box (1978) y Armitage (1996) son ejemplos muy conocidos. También Sokal (1969), Zar (1999), Colton (1979), Ching Chu Li (1969), Milton (2001) y Rothman (1998) han seguido semejantes criterios docentes, y en España los magníficos textos de Romero y Zúnica (1986), Doménech (1989), y Martín y Luna (1994) están en la misma línea.

3. No hay en español libros con este enfoque

Aunque con cierta frecuencia se publican libros y artículos explicando al lector no versado en Estadística lo que el valor P del test indica, son mayoría las publicaciones que enseñan los tests estadísticos proponiendo el uso de la cantidad 0.05 como referencia. Esa propuesta contiene matices que suelen escapar al lector sin conocimientos estadísticos, de modo que interpreta que esa cantidad tiene propiedades intrínsecas de frontera que separa los resultados válidos de los inservibles.

Este libro explica al investigador sin conocimientos matemáticos los fundamentos de la Inferencia Estadística, de modo que pueda entenderla y ejecutarla apropiadamente y sintiéndose seguro de lo que hace. Muestra que la *investigación científica* y la *Toma de Decisiones* son dos procesos diferentes que requieren diferente estrategia de Inferencia. Analiza las causas del auge de la «regla del 5%» y pretende colaborar eficazmente a evitar la confusión y contradicciones que se siguen de la aplicación indiscriminada de esa desafortunada “receta”. En castellano es escasa la literatura que denuncia estos errores y explica el modo de evitarlos.

¿Quién no conoce la siguiente copla, sin duda inspirada a los investigadores por su tormentosa relación con la Inferencia Estadística?

*Ni contigo ni sin ti
tienen mis males remedio,
contigo porque me matas,
y sin ti porque me muero.*

Esperamos que este libro contribuya «significativamente» a la urgente tarea de deshacer un malentendido, casi centenario, y liberar a los investigadores de un tabú que entorpece su buen hacer.

NOTA: Antes de continuar la lectura del libro le sugerimos realice las encuestas de autoevaluación del Apéndice 1, diseñadas para ayudarle a valorar su nivel de conocimientos previos en estos temas.

Si responde correctamente a la inmensa mayoría de las preguntas, le resultará curioso leer los capítulos 2 y 3, en los que se enuncian y documentan los errores más frecuentemente cometidos por la mayoría de los investigadores. Usted no necesitaría leer del capítulo 4 en adelante donde se analiza detalladamente cada uno de estos errores y se explica el tema correspondiente.

Si usted no respondió correctamente a la inmensa mayoría de las preguntas, le animamos a que lea este libro y compruebe después que ya no comete esos errores volviendo a hacer estas encuestas.

BIBLIOGRAFÍA

- Armitage P. y Berry G. «Statistical methods for medical researchers». Blackwell. 1996.
- Box, G.E. y otros. «The Design and Analysis of Industrial Experiments». Longman. 1978.
- Colton, T. «Estadística en Medicina». Salvat. 1979.
- Ching Chung Li. «Introducción a la Estadística Experimental». Omega. 1969.
- Cochran, W.G. y Cox, M.G. «Experimental Designs». John Wiley. 1957.
- Doménech, J.M. «Bioestadística». Herder. 1989.

- Feinstein A.R. «Clinical epidemiology: the architecture of clinical research». Philadelphia, WB saunders. 1985. Citado por F.L. Redondo. «El error en las pruebas de diagnóstico clínico». Díaz de Santos. 2002. (p. 52).
- Fisher R.A. «Statistical methods for research workers». Hafner Press. 1925.
- Fisher R.A. «The design of experiments». Hafner Press. 1935.
- Fisher R.A. «Statistical methods and scientific inference». Hafner Press. 1956.
- Martín Andrés, A. Y Luna del Castillo, J. «Bioestadística para Ciencias de la Salud». Norma. 1994.
- Milton, J.S. «Estadística para Biología y Ciencias de la Salud». Interamericana. 2001.
- Moyé L.A. «Statistical Reasoning in Medicine. The intuitive P-Value Pimer». Springer. 2000.
- Romero, R. y Zúñiga, L. Estadística. Universidad de Valencia. 1986.
- Rothman K. y Greenland W. «Modern Epidemiology». Lippincott-Raven Pub. 1998.
- Rothman K. «Modern Epidemiology». Little Brown. Toronto. 1986.
- Scheffler, W.C. «Statistics for health professionals». Adisson-Wesley. 1984.
- Snedecor G. y Cochran W.G. «Statistical Methods». John Wiley and Sons. 1950.
- Sokal, R.R. y Rohlf, F.J. «Biometry». Freeman and Co. 1969.
- Winner, B.J. «Statistical principles in experimental design». McGraw-Hill. 1970.
- Yule, J.U. y Kendall. M.G. «An introduction to the theory of Statistics». Griffin. 1953.
- Zar, J.H. «Biostatistical Analysis». Prentice Hall. 1999.

El testimonio de los expertos

Para el investigador acostumbrado a expresar las conclusiones de sus trabajos con frases aparentemente contundentes como «significativo» o «no significativo» es difícil renunciar a ese esquematismo y aceptar que esos términos no tienen utilidad real en el contexto de la investigación científica. Podría pensar que los argumentos expuestos en este libro expresan una opinión muy particular de los autores, pero la realidad es que reflejan el pensamiento de muchos expertos en análisis de datos. Continuamente aparecen libros y artículos explicando que ni el valor 5% ni otro convenido al respecto tienen carácter de frontera conceptual que separe los resultados útiles de los inservibles.

Aunque en los capítulos que lo requieren se dan las citas pertinentes, en éste se recogen las que ponen de relieve la contundencia de su postura. Todos ellos insisten en que en el proceso de formarse opinión nuestra mente no usa puntos de separación dicotómica, sino escalas de variación continua.

Siguiendo un orden cronológico comenzaremos citando a:

Snedecor y Cochran (1960): «Debe evitarse esa actitud que consiste en considerar los tests de significación como una regla para decidir de modo automático si se acepta o se rechaza una hipótesis. El uso del 5% o el 1% es simplemente una convención. Es loable la práctica seguida por algunos autores de publicar el valor P encontrado al hacer el test».

Box (1982): «Es mejor reportar el valor de P encontrado que decir si es o no «significativo» a un nivel convenido. La afirmación de que un

resultado «no fue estadísticamente significativo al nivel de 5%» quiere decir en muchas ocasiones que el valor de P fue del orden de 0.06. Y la diferencia de actitud mental asociada a 0.05 y a 0.06 es despreciable, naturalmente. Los tests estadísticos han sido sobre-utilizados y en muchos de los casos en que se usan hubiera sido mas informativo dar el intervalo de confianza para el valor poblacional del parámetro».

Rothman (1986), citando un trabajo de Freiman (1978) publicado en el *New England Journal of Medicine*: «En una revisión de 71 Ensayos Clínicos reportados con diferencia «no significativa» entre los tratamientos comparados, se encontró que en la mayoría de ellos los datos eran compatibles con un efecto moderado o incluso razonablemente fuerte del tratamiento nuevo. En todos estos casos los investigadores interpretaron que no había efecto porque el valor P no fue «estadísticamente significativo». No pudiendo rechazar la hipótesis nula el investigador deduce inapropiadamente que es cierta, lo cual probablemente es falso en muchos de los estudios con resultado «negativo». El inadecuado proceder de los autores en estos casos se habría evitado si se hubieran concentrado en los intervalos de confianza, más que en los tests».

Bourke, Daly and McGilvray (1985): «La receta del 5% es universalmente usada, pero claramente arbitraria... no tiene ningún sentido rechazar una hipótesis nula porque es $P = 0.0499$ y no rechazarla si es $P = 0.0501$ ».

Armitage (1988): «El valor 5% ha llegado a ser ampliamente usado como un punto de corte para decidir la relevancia del alejamiento de los datos respecto a lo que propone la hipótesis nula. Esto es desafortunado, porque no debe haber distinción rígida entre un valor P justamente por debajo del 5% y otro que está escasamente por encima. Es preferible evitar la dicotomía «significativo» y «no significativo» y decir cuan acusado es el alejamiento respecto a los valores esperados dando el valor P ».

«Además hay que tener muy claro que aunque un valor pequeño de P constituye evidencia contra la hipótesis nula, un valor grande no constituye evidencia a favor».

Rothman (1998): «Cuando un solo estudio constituye la única base para elegir entre dos posibles acciones, como en situaciones de control de calidad en la industria, la metodología de la Toma de Decisiones es la adecuada. Pero en la mayoría de las investigaciones es presuntuoso, sino absurdo, que el investigador actúe como si su trabajo fuera la única fuente para tomar decisiones. Las decisiones se toman inevitablemente a par-

tir de varios estudios y el razonamiento correcto requiere mucho más que la clasificación de un estudio en «significativo» o «no significativo». Esa degradación de la información acerca de un efecto en una simple dicotomía es contraproducente y puede conducir a errores» (Pág.187).

«¿Por qué tan infundada simplificación dicotómica se hizo tan popular en la investigación científica? Indudablemente, mucha de la popularidad de estos métodos procede de su aparente objetividad y la rotundidad de las expresiones usadas. El declarar un efecto como «significativo» o «no significativo» oculta la necesidad de un razonamiento prudente. Esos calificativos sirven como un sustituto mecánico del análisis lógico. La presunta nitidez de las conclusiones enunciadas es más gratificante para investigadores, editores y lectores que una conclusión menos cuadrículada. Ya que los tests estadísticos son tan frecuentemente malinterpretados, recomendamos evitar la clasificación dicotómica en «significativo» y «no significativo» (Pág. 194).

Sterne and Smith (2001): «Un problema fundamental es la general incomprensión de la naturaleza de los tests estadísticos. El investigador debe tener muy presentes estos puntos:

1. Debe dejar de creer que el valor 5% tiene una especial importancia.
2. La arbitraria división de los resultados en «significativos» y «no significativos» (de acuerdo a la frontera del 5% comúnmente usada) no fue la intención de los fundadores de la Inferencia Estadística. Fisher vio el valor P como un índice que mide la fuerza de la evidencia contra la hipótesis nula.
3. La calificación de diferencias como «estadísticamente significativas» no es aceptable. En la sección de resultados debe ser presentada la fuerza de la evidencia contra la hipótesis nula, es decir, el valor P, sin referencia a un umbral arbitrario.
4. En muchos casos la publicación de resultados en investigación médica no requiere tomar decisiones. Esos resultados contribuyen a incrementar el cuerpo de conocimientos sobre el tema que se investiga».

En carta al editor aparecida en la misma revista y con la misma fecha D. R. Cox, quizá la voz más respetada actualmente en el ámbito de la Estadística, muestra su acuerdo con estos criterios expuestos por Sterne y Smith.

Romero Villafranca (2004):

1. La utilización de valores límites o «críticos» para el p-value (como el 5% o el 1%), que separan los resultados «significativos» de los «no significativos» es un anacronismo, reflejo de épocas en las que el investigador no tenía acceso al cálculo exacto de estos valores y debía referirse a tablas que se limitaban a estos dos niveles. Es el p-value lo que refleja el grado de evidencia de unos resultados contra la H_0 y, en consecuencia, lo que debería acompañar al análisis de los datos, y no sólo la constatación de si resulta superior o inferior al 5%. ¿Qué diferencia hay, en la práctica, entre un p-value del 4.9% o del 5.1%?
2. Otro error muy frecuente es la confusión entre significación estadística e importancia práctica. Si cierta diferencia es «muy significativa estadísticamente», la interpretación correcta es que es casi seguro que dicha diferencia no es nula, y no necesariamente que sea muy importante. En este sentido, el cálculo del *intervalo de confianza* para el efecto en cuestión es mucho más informativo que la simple constatación de si dicho intervalo contiene o no al cero, que en el fondo es lo que hace el test de hipótesis.
3. Que un test estadístico no rechaza una H_0 , no significa que quede demostrado que dicha hipótesis es cierta, sino sólo que es compatible con los datos observados, como lo son probablemente también muchas otras hipótesis. Nuevamente el *intervalo de confianza* es más informativo, a efectos de ayudar a tomar posición.
4. En el campo de la investigación científica, que unos resultados no lleguen a ser significativos estadísticamente (entendido ello de la forma habitual, como que el p-value sea superior al 5%) no significa necesariamente que no merezcan ser publicados, especialmente si los efectos constatados van en el sentido que sugieren las hipótesis de trabajo de la investigación. Esos resultados, acumulados con otros sobre el tema, pueden permitir llegar a la comunidad científica a conclusiones sólidas.
5. El análisis gráfico de los «residuos» debería ser una práctica ineludible en cualquier estudio estadístico, y las revistas científicas deberían ser más exigentes al respecto, en vez de la preocupación obsesiva que algunas muestran por el mítico 5%.

Pero aunque en el quehacer científico es importante el criterio de autoridad, aun lo es más la fuerza de la razón. Los criterios defendidos y explicados en este libro tiene su mayor defensa en el sentido común. El lector es invitado a formarse su opinión y usar los tests estadísticos con sensatez y eficacia, y sin depositar en ellos expectativas desmesuradas. Téngase muy presente que no hay sutiles y arcanas «razones matemáticas» que se opongan al sentido común. Recordemos con Bertrand Russel que la matemática es básicamente la formalización de la lógica, lo que la hace más potente pero no contraria a ella.

BIBLIOGRAFÍA

- Armitage P. y Berry G. «Statistical methods for medical researchers». Blackwell. 1996. pp. 95 y 96.
- Bourke D.D., Daly L.E. y McGilvray J. «Interpreting and use of medical statistics». Blackwell. 1985. 3ª edit. p. 71.
- Box, G.E., Hunter, W.G. y Hunter, J. S. «Statistics for Experimenters» John Wiley. 1982. Cap. 5, p. 109.
- Freiman, J.A., Chalmers, T.C., Smith, H. *et al.* «The importance of beta, the Type II error and sample size in the design and interpretation of the randomised control trial. Survey of 71 «negative» trials. N. England. J. Med. 1978; 299: pp. 690-694.
- Romero Villafranca, R. Ver apéndice 4 de este libro. 2004.
- Rothman K. «Modern Epidemiology». Little Brown. Toronto. 1986. pp. 187 y 193.
- Rothman K. y Greenland W. «Modern Epidemiology». Lippincott-Raven Pub. 1998. pp. 187 y 194.
- Rothman K.J. «A show of confidence». New England of Journal of Med. 299: pp. 1362-1363. 1978.
- Snedecor G. y Cochran W.G. «Statistical Methods» . John Wiley and Sons. 1960. Cap. 1.
- Sterne J.A.C. y Smith G.D. «Sifting the evidence. What's wrong with significance tests?» BMJ vol. 322. pp. 226-231. 2001.

Los errores más graves y frecuentes

Como hemos apuntado antes, creyendo que para entender el significado del valor P de los tests se requieren unos conocimientos matemáticos de los que ellos carecen, la mayoría de los investigadores biológicos renuncian a comprender ese tema. Pero como tienen que usarlo inexcusablemente optan por aferrarse a unas reglas rígidas, que al ser usadas sin discernimiento, pueden generar más confusión que ayuda.

En este capítulo mostramos, a través de un ejemplo concreto, los errores que más frecuentemente se cometen. A lo largo de la obra se ven detenidamente cada uno de estos errores y el modo de evitarlos.

Para estudiar el posible efecto anticancerígeno (AC) de 4 productos recientemente descubiertos, «A», «B», «C» y «D», trabajaremos con ratas de una cepa genéticamente modificada en la que el 60% de ellas desarrollan cáncer de cérvix espontáneamente.

Probaremos cada fármaco en 40 ratas. Para cada uno de ellos, si no es AC esperamos que unas 24 hagan cáncer (24 es el 60% de 40). Cuanto menor sea el número de ratas que desarrollan cáncer más nos inclinaremos a pensar que hay efecto AC. Cuando el resultado no es concluyente se calcula el «valor P del test» y es entonces cuando pueden cometerse gruesos errores de interpretación.

He aquí los resultados obtenidos y el valor P de los tests estadísticos¹:

¹ Los valores P de esta tabla son unilaterales, como lo son también los del resto de la obra, mientras «no se diga lo contrario».

Fármaco	Núm. de ratas con cáncer en la muestra	% de ratas con cáncer en la muestra	Valor P
A	8	20%	0,000003
B	18	45%	0,039
C	19	47,5%	0,074
D	23	57,5%	0,436

Muchos investigadores enunciarían las conclusiones con esta frase:

«Habiendo decidido declarar como significativos los efectos con $P < 0,05$ concluimos que “A” y “B” son anticancerígenos ($P < 0,05$), mientras que “C” y “D” no lo son ($P > 0,05$)».

Y la mayoría de sus lectores darían por buena esa expresión. Sin embargo en ella se encierran hasta 6 errores serios.

1. Establece conclusiones muy diferentes para «B» y «C», cuando en realidad es muy similar lo que se puede concluir acerca de uno y otro.

En realidad los datos tienen el mismo valor si se encuentra P ligeramente superior o ligeramente inferior al 5%. Es decir, $P = 0,074$ y $P = 0,039$, por ejemplo, nos dicen prácticamente lo mismo.

En general, un error muy frecuente es no publicar o infravalorar un resultado calificándolo como «no significativo» por el hecho de que la P sea un poco mayor del 5%.

2. Establece conclusiones semejantes para «A» y «B», cuando en realidad es muy distinto lo que se puede concluir acerca de uno y otro.

En general, es un error frecuente reportar como « $P < 0,05$ » el resultado de un test tanto si se obtuvo realmente, por ejemplo, $P = 0,000003$, como si se obtuvo $P = 0,039$.

Con un valor de $P = 0,039$ la evidencia a favor de que el tipo de efecto encontrado en la muestra es una realidad en la población general es muy modesta, mientras que con $P = 0,000003$ esa evidencia sería prácticamente definitiva.

3. Los resultados muestran que «D» *puede ser* inútil, lo cual no debe confundirse con que permiten afirmar que lo es.

En general, es un error frecuente confundir, con valores de P grandes, «La hipótesis *puede ser* cierta» con «La hipótesis *es* cierta».

En realidad un valor grande de P nos dice que pueden ser ciertas muchas hipótesis. Asegurar que la cierta es precisamente una de ellas suele ser totalmente injustificado (la ausencia de evidencia no implica evidencia de ausencia).

4. Estos resultados ofrecen evidencia clara a favor de que «A» es AC, pero para los otros tres fármacos no se puede tomar postura porque los datos son compatibles con que sean AC, pero también lo son con que no sean AC.

En general, es un grave error ignorar las limitaciones propias de los tests de significación, forzando la elaboración de conclusiones nítidas (asegurar que existe o no realmente cierto efecto) en toda investigación.

La realidad es que en muchos estudios es imposible pronunciarse definitivamente a favor o en contra de una hipótesis. Decir que los resultados son «significativos» o «no significativos» no disminuye el nivel de incertidumbre propio de esos resultados.

5. Frases como «se decide considerar significativos los resultados con $P < 0,05$ » no tienen sentido en la investigación científica, en la que no se trata de decidir, sino de valorar en qué medida los resultados obtenidos constituyen evidencia a favor de cierta hipótesis.

En general, es un craso error confundir la finalidad de la investigación científica, destinada a valorar en qué medida los datos aportan evidencia a favor de cierta hipótesis, con la finalidad de la *Toma de Decisiones*, en la que a partir de lo observado en la muestra se decide una u otra acción.

«... el propósito central de un experimento no es precipitar la toma de decisiones sino propiciar un reajuste en el grado de confianza que uno tiene en la veracidad de cierta hipótesis... la tarea del científico no es prescribir acciones, sino establecer convicciones razonables» Silva (1997)².

² Silva LC. *Cultura estadística e investigación científica en el campo de la salud*. Díaz de Santos, 1997.

6. No se dan los intervalos de confianza para el efecto real de cada uno de los productos.

En general, el no comentar la magnitud del efecto encontrado en la muestra y del intervalo de confianza correspondiente dificulta notablemente la elaboración de conclusiones correctas.

Lo correcto es resumir los resultados obtenidos informando no solo del valor P del test, sino también dando el intervalo de confianza para el % poblacional de cánceres con cada uno de los fármacos, en cuyo caso, no hay que caer en el error de atribuir a los límites de dicho intervalo carácter de frontera determinante (como no lo es tampoco un posible valor de P convenido por el investigador).

He aquí los resultados, el valor P del test y los intervalos de confianza:

Sin ningún tratamiento hacen cáncer el 60% de las ratas.

Se da cada fármaco a 40 ratas. El 60% de 40 es 24

Fármaco	Núm. de ratas con cáncer	% de ratas con cáncer	Valor P IC al 99%
A	8	20%	0,000003 7%-41%
B	18	45%	0,039 25%-66%
C	19	47,5%	0,074 27%-68%
D	23	57,5%	0,436 36%-77%

A la vista de esos intervalos de confianza las conclusiones razonables son:

«El fármaco “A” parece ser un potente AC que probablemente baja la incidencia de cáncer entre 53 y 19 puntos, mientras que “B, C, D” pueden serlo o no serlo»³.

³ «B, C y D» puede que bajen la incidencia en 35, 33 y 24 puntos respectivamente (o algo más), pero también puede que no la modifiquen e incluso que la incrementen.

Vea la notable diferencia entre esta conclusión y la enunciada inicialmente, asumiendo que 0,05 es una barrera definitiva:

«Habiendo decidido declarar como significativos los efectos con $P < 0,05$ concluimos que “A” y “B” son anticancerígenos ($P < 0,05$), mientras que “C” y “D” no lo son ($P > 0,05$)».

Las dos conclusiones solamente coinciden para el fármaco «A», y difieren notoriamente para los otros tres productos.

En el siguiente capítulo comenzamos a explicar la lógica de los tests de significación y el uso correcto del valor P del test.

La inferencia en la vida común

Todas las encuestas muestran que entender lo que indica el valor «P» de los tests es el mayor problema de los investigadores biológicos, tanto al elaborar las conclusiones de sus propias investigaciones como al interpretar los resultados publicados por sus colegas.

En este y los siguientes capítulos se explica este tema sin utilizar herramienta matemática alguna y de modo que toda persona interesada pueda, tras una lectura detenida y cuidadosa, entender el concepto básico y su aplicación práctica.

Empezando por reflexiones muy sencillas iremos acercándonos paulatinamente al meollo de la cuestión hasta llegar cómodamente a su centro. No queremos llegar a la cima escalando pendientes escarpadas, sino siguiendo un sendero amigable que rodea la montaña.

Leyendo con atención lo que sigue, todo profesional llega a entender claramente el tema de la P del test. Esa cierta inseguridad e incomodidad que muchos investigadores sienten cuando se les cruza este tema en el camino serán sustituidas por confianza y seguridad.

TESTS DE SIGNIFICACIÓN, UN PROCESO COMÚN Y SENCILLO

Lo primero que debe quedar claro es que el tipo de razonamiento propio de los tests de significación (TS) es muy sencillo y lo usamos en la vida cotidiana constantemente. No solo las personas con nivel cultural

universitario, sino las de cualquier nivel cultural y los niños a partir de los siete años aproximadamente.

Llama poderosamente la atención que, siendo un mecanismo mental, sencillo y común a todos los humanos, muchos investigadores lo consideren un proceso de alto contenido matemático que ellos renuncian a entender. Esta errónea interpretación de los TS es consecuencia de la deficiente enseñanza que han recibido en este campo. Siendo el Análisis Estadístico una disciplina esencialmente matemática, los profesores tienden a explicarla usando un lenguaje matemático ininteligible para los profesionales de otras ciencias.

De ese modo se encuentran atrapados entre la incomprensión de este proceso y la obligación de usarlo por imperativo de la comunidad científica. La salida mayoritariamente elegida a esta situación es aferrarse con fe ciega a la repetición de ciertas muletillas que tienen su origen en ideas sensatas, pero al ser usadas fuera de contexto pierden toda su utilidad y cuyo contenido no entiende realmente ni quien las escribe ni quien las lee.

Aquí se exponen los fundamentos lógicos de los TS, de modo que todo lector pueda entenderlos sin ambigüedad. Se empieza poniendo ejemplos de la vida común en los que usamos ese mismo proceso mental. Y se continúa mostrando que al elaborar las conclusiones de los trabajos científicos se usa el mismo proceso lógico.

Intente meterse en los ejemplos que siguen usando simplemente su sentido común. Le pueden parecer excesivamente sencillos, obvios y sin relación con la elaboración de conclusiones en la investigación biomédica. Pero muy pronto usted verá que el elemental proceso lógico usado en estas situaciones de la vida común es el mismo que se usa en la Inferencia Estadística para elaborar las conclusiones de los trabajos científicos.

LOS TESTS DE SIGNIFICACIÓN EN LA VIDA COMÚN

Se dice habitualmente que la ciencia avanza planteando hipótesis y haciendo observaciones o experimentos que nos permitan decidir si son ciertas o falsas. Pero Popper, en la línea de pensamiento iniciada por Hume, hizo notar que muy raramente los experimentos permiten confirmar que una hipótesis es cierta. Los resultados experimentales solo permiten una de estas dos cosas:

- a) Si son incompatibles con la hipótesis nos llevan a concluir que es *falsa*, la rechazamos.
- b) Si son compatibles con la hipótesis nos llevan a concluir que *puede ser cierta*, pero no nos aseguran que lo sea, porque esos resultados son también compatibles con otras hipótesis.

Veamos algunos ejemplos muy sencillos:

1. *¿Hay o hubo materia orgánica en Marte?*

Considere la hipótesis: en Marte no hay ni hubo nunca materia orgánica.

Nuestra observación: analizar tres muestras de suelo marciano traídas por una sonda espacial.

- a) Si en esas muestras descubrimos algún rastro de materia orgánica rechazamos la hipótesis.
- b) Si no encontramos ningún rastro de materia orgánica, decimos que nuestros datos son compatibles con la hipótesis. La hipótesis *puede* ser cierta, pero no afirmamos que lo sea. De momento la mantenemos como válida, mientras un nuevo dato no nos obligue a rechazarla.

2. En una mansión señorial se comete un crimen a las 12:00 horas y Bautista, el mayordomo, es uno de los sospechosos.

Considere la hipótesis, H_0 : *Bautista es el asesino*.

Nadie ha visto cometer el asesinato, pero algunos testigos pueden dar información inequívoca sobre la ubicación de Bautista a las 12:15.

¿Qué decisión tomaríamos en cada uno de estos casos y por qué?

1. Bautista fue visto a las 12:15 h a 900 km de la casa. Concluimos que:

- a) Bautista *no* es el asesino (rechazo H_0).
- b) Bautista *puede* ser el asesino (acepto H_0 como posible).
- c) Bautista *es* el asesino (afirmo que H_0 es cierta).

2. Bautista fue visto a las 12:15 h en el portal de la mansión.

Concluimos que:

- a) Bautista *no es* el asesino (rechazo H_0).
- b) Bautista *puede* ser el asesino (acepto H_0 como posible).
- c) Bautista *es* el asesino (afirmo que H_0 es cierta).

3. Bautista fue visto a las 12:15 en la habitación contigua a la del crimen. Concluimos que:

- a) Bautista *no es* el asesino (rechazo H_0).
- b) Bautista *puede* ser el asesino (acepto H_0 como posible).
- c) Bautista *es* el asesino (afirmo que H_0 es cierta).

Si en el primer caso usted ha elegido la opción «a» y en el segundo y tercero la «b» estará de acuerdo con el resto de los mortales. Si eligió otras opciones, reflexione de nuevo sobre este tema.

Especialmente erróneo sería decir que Bautista es el asesino en los casos segundo y tercero. Y veremos que la mayoría de los investigadores cometen, al hacer la Inferencia Estadística, un error conceptual equivalente a decir que Bautista *es* el asesino, en vez de decir que *podría serlo*. El hecho es muy llamativo porque este atentado flagrante contra el sentido común lo perpetran tanto los investigadores inexpertos como los más cualificados.

Felizmente, evitar tan graves equivocaciones no es una cuestión de matemáticas. Basta con que el lector sea consciente de que se trata de hacer lo mismo que hacemos en la vida cotidiana.

Reflexione sobre el tipo de razonamiento que hemos hecho en estos supuestos:

1. Rechazamos la hipótesis planteada cuando la información de que disponemos es incompatible con ella, es decir, si fuera cierta la hipótesis sería muy difícil que se diera el hecho que se ha dado.
2. Cuando el hecho observado es compatible con la hipótesis, no afirmamos que sea cierta (no decimos que Bautista es culpable). Solo decimos que *puede ser* cierta, pues el dato no constituye evidencia contra ella.

LA HIPÓTESIS NULA, H_0 , PUNTO DE PARTIDA OBLIGADO

Veamos otros ejemplos en la misma línea, para introducir el concepto de «hipótesis nula», fundamental para entender el proceso lógico de los tests de significación.

Consideremos el caso de un centinela que tenía encomendada la vigilancia y protección de una mansión y ciertos indicios sugieren que quizá abandonó la guardia, incurriendo en gravísimo delito. Al iniciarse el juicio el juez hace saber que en principio todo ciudadano es inocente y solo se le declarará culpable si se aportan datos que lo prueben.

Esta postura tan lógica desde el punto de vista jurídico tiene un estrecho paralelismo con la postura de la comunidad científica al evaluar los descubrimientos presentados por los investigadores. El paralelismo es tan claro y ayuda tanto a entender la estructura lógica de los tests de significación, que merece la pena insistir en él.

En *derecho penal*: hay que evitar condenar a los inocentes. Por ello partimos de la *hipótesis inicial de inocencia* y solo consideraremos culpable al acusado si hay pruebas claras contra él, si hay datos claramente incompatibles con la hipótesis de inocencia.

En *investigación científica*: hay que evitar elevar a la categoría de «hechos biológicos generales» hallazgos que sean solo una «anécdota de la muestra estudiada». Por ello partimos de la *hipótesis que dice que el hallazgo presentado por un investigador podría ser una mera anécdota de la muestra estudiada y no una realidad de validez general*. A esta hipótesis se la llama «hipótesis nula» y se la representa por H_0 , y la rechazaremos —considerando que ese hallazgo es una realidad de validez general— cuando los datos aportados sean incompatibles o difícilmente compatibles con ella.

Por ejemplo, si los resultados de un investigador sugieren que cierto producto es cancerígeno, la postura de la comunidad científica es no aceptar esa «acusación de culpabilidad» mientras no haya hechos que la avalan claramente. Es decir, se parte de la hipótesis nula de «inocencia» de ese producto y solo se le considerará cancerígeno si los datos presentados por el investigador son incompatibles con la inocencia.

Pero en la vida común tampoco la sociedad quiere premiar con honores a un ciudadano si no hay motivos reales para ello. Si los amigos de un aspirante a un premio dicen que esa persona es excepcionalmente buena,

la comisión encargada de valorar el caso parte de una postura inicial más bien escéptica, en el sentido de considerar que toda persona es básicamente normal hasta que no se demuestre que ha hecho méritos merecedores de honores especiales.

De modo equivalente, cuando un científico presenta resultados a favor de que cierto producto ayuda a prevenir o curar cierta enfermedad, la comunidad científica adopta una postura inicial escéptica, que no consiste en negar lo que dice el investigador, pero sí en no aceptar que hay efecto curativo mientras no se aporten datos claros en ese sentido. Por ello se plantea la hipótesis nula, H_0 , que dice que *ese producto no es curativo*, y solamente se abandona esa postura y se asume que es beneficioso cuando los resultados presentados son incompatibles con que sea inútil.

Para aplicar correctamente un TS lo primero es tener claro cuál es la hipótesis nula planteada en ese caso y la experiencia muestra que muchos investigadores cometen errores de interpretación del TS precisamente porque no identifican cuál es la H_0 en el caso particular de su investigación.

En el caso de un presunto culpable, la H_0 establece que no lo es. En el caso de presunto héroe, la H_0 establece que no lo es. En el caso de una sustancia presuntamente perjudicial para la salud, la H_0 establece que realmente es inocua. En el caso de un fármaco presuntamente beneficioso, la H_0 establece que realmente es inútil. En general, la H_0 establece que en la población general no hay el tipo de efecto encontrado en la muestra, sino que ese efecto ha aparecido en la muestra por casualidad.

RECHAZAR O NO RECHAZAR LA H_0 , HE AHÍ LA CUESTIÓN

Consideremos el ejemplo de los centinelas Abel, Blas y Caín, que tenían encomendada la protección de la mansión a las 5 de la tarde y ciertos indicios sugieren que quizá alguno de ellos abandonó la guardia. No hay testimonio fiable sobre dónde se encontraba cada uno a las 5, pero los hay acerca de dónde se encontraban a las 6 de la tarde.

¿Qué decisión tomamos en cada uno de estos casos y por qué?

1. En el juicio contra Abel se parte de la hipótesis nula, H_0 , que dice que es inocente, es decir, que a las 17:00 h estaba en la casa.

Hechos observados: Abel fue visto a las 18:00 h a 5.000 km de la casa.

A la vista de ese dato concluimos:

- a) Rechazamos $H_0 \rightarrow$ *No estaba* en la casa a las 17:00 h.
 - b) Aceptamos H_0 como posible $\rightarrow$ *Puede que estuviera* en la casa a las 17:00 h.
 - c) Afirmamos que H_0 es cierta $\rightarrow$ *Estaba* en la casa a las 17:00 h.
2. En el juicio contra Blas se parte de la hipótesis nula, H_0 , que dice que es inocente, es decir, que a las 17:00 h estaba en la casa.

Hechos observados: Blas fue visto a las 18:00 h a 50 km de la casa.

A la vista de ese dato concluimos:

- a) Rechazamos $H_0 \rightarrow$ *No estaba* en la casa a las 17:00 h
 - b) Aceptamos H_0 como posible $\rightarrow$ *Puede que estuviera* en la casa a las 17:00 h.
 - c) Afirmamos que H_0 es cierta $\rightarrow$ *Estaba* en la casa a las 17:00 h.
3. En el juicio contra Caín se parte de la hipótesis nula, H_0 , que dice que es inocente, es decir, que a las 17:00 h estaba en la casa.

Hechos observados: Caín fue visto a las 18:00 h en la puerta de la casa.

A la vista de ese dato concluimos:

- a) Rechazamos $H_0 \rightarrow$ *No estaba* en la casa a las 17:00 h.
- b) Aceptamos H_0 como posible $\rightarrow$ *Puede que estuviera* en la casa a las 17:00 h.
- c) Afirmamos que H_0 es cierta $\rightarrow$ *Estaba* en la casa a las 17:00 h.

Si usted ha elegido «a» en el caso de Abel y «b» para los otros, estará de acuerdo con los demás lectores.

Para hacer la inferencia en la vida común:

Formulamos una hipótesis y observamos unos hechos.

- Si los hechos son difícilmente compatibles con la hipótesis, la rechazamos.
- Si los hechos son compatibles con la hipótesis decimos que puede ser cierta, pero no aseguramos que lo sea.

TESTS DE SIGNIFICACIÓN E INTERVALOS DE CONFIANZA

Al hacer inferencia, tanto en la vida común como en la actividad científica, puede ser procedente calcular los llamados «intervalos de confianza». Veámoslo en un ejemplo muy sencillo no científico.

Estamos interesados en saber la edad de una persona, disponiendo como única información de algunas fotos recientes. Es obvio que por la imagen de la fotografía no podemos decir exactamente la edad que tiene. Pero sí podemos saber que *no tiene* ciertas edades que son claramente incompatibles con la imagen. Nos planteamos una hipótesis sobre su edad y la rechazamos o no según que su apariencia externa sea incompatible o compatible con esa edad.

Además, hay otro modo de referirnos a esa edad que desconocemos, diciendo que presumiblemente estará comprendida entre dos cantidades, es decir, dar lo que se llama un «intervalo de confianza», IC, dentro del cual creemos que está la edad de esa persona.

La ayuda de los TS y los IC al elaborar las conclusiones de una investigación biomédica implica exactamente el mismo razonamiento que en estas situaciones de la vida común. Insistamos en este ejemplo

para mostrar que el cálculo de los intervalos de confianza es un complemento necesario de los tests de significación. Veamos esta fotografía actual de Ana Belén.

Supongamos que estamos interesados en saber su edad y nos planteemos las siguientes hipótesis:

- a) Ana Belén tiene **2** años. → Rechazamos esta hipótesis porque el dato, en este caso la fotografía, es incompatible con ella.
- b) Ana Belén tiene **14** años. → Rechazamos esta hipótesis porque el dato es incompatible con ella.



FOTÓGRAFO: JAVIER PELLICER VIDAL

- c) Ana Belén tiene 7 años. → No rechazamos esta hipótesis porque el dato, en este caso la fotografía, es compatible con ella.
- d) Ana Belén tiene 8 años. → No rechazamos esta hipótesis porque el dato es compatible con ella.

Otro modo de hablar sobre la edad que Ana Belén puede tener es dar un «intervalo de confianza». Podríamos decir que su edad probablemente está *entre 6 y 9 años*, lo que es coherente con haber rechazado que tenga 2 o 14 y haber aceptado que puede tener 7 u 8.

Resumiendo

- a) **Test de significación:** es el proceso lógico que nos lleva a descartar una hipótesis si encontramos datos incompatibles con ella, o a aceptarla como posible si los datos son compatibles con ella. Se plantea la hipótesis que dice que Ana Belén tiene determinada edad, y a la vista de la información aportada por la fotografía concluimos que *no tiene* esa edad o que *puede tenerla*.
- b) **Intervalo de confianza:** a partir de la información parcial de que disponemos, podemos calcular un intervalo dentro del cual probablemente estará el valor que querríamos conocer. El modo más razonable de resumir lo que podemos decir acerca de esa edad es dar el IC dentro del cual creemos que esté comprendida. No sabemos exactamente la edad de Ana Belén, pero pensamos que estará entre dichos valores.

En el siguiente capítulo veremos que al hacer Inferencia Estadística, para elaborar conclusiones razonables en investigación biológica se procede de modo totalmente paralelo al de estos ejemplos.

COMPRUEBE SU NIVEL DE CONOCIMIENTOS EN ESTE TEMA

En el Apéndice 2 encontrará una encuesta de autoevaluación para este capítulo, que le ayudará a evaluar en qué medida tiene claras sus ideas en este tema.

Capítulo 5

La Inferencia Estadística en investigación médica

Se veía en el capítulo anterior cómo hacemos los tests de significación (TS) y calculamos los intervalos de confianza (IC) en la vida común.

En este se ve que los TS y los IC usados para elaborar conclusiones razonables en la investigación implican exactamente el mismo proceso lógico, y por tanto se entienden sin ningún problema.

Antes de ver ejemplos concretos de TS e IC en la investigación biomédica, se revisan algunas ideas fundamentales sobre la investigación cuantitativa en biomedicina.

MUCHAS AFIRMACIONES BIOLÓGICAS SE REFIEREN A MEDIAS Y PROPORCIONES POBLACIONALES

Hay que ser conscientes de que muchas de las afirmaciones que se hacen en biología tienen, aunque a primera vista no lo parezca, un contenido estadístico, ya que aluden al valor de la *media* de una magnitud en una *población* de individuos o a la *proporción* de ellos que tiene cierta característica. Puesto que la proporción puede considerarse como un caso particular de media, en adelante diremos solamente «media» en vez de «media o proporción».

Por ejemplo, al afirmar que *los recién nacidos (RN) varones pesan más que las hembras*, no queremos decir que todo varón pese más que cualquier hembra. De hecho, se pueden encontrar numerosos ejemplos particulares de un varón que pesa menos que una hembra. Pero esos

casos particulares no invalidan nuestra afirmación, puesto que ella se refiere a los *valores medios en las poblaciones*. Concretamente, quiere decir que el peso *medio* de la *población* de los RN varones es mayor que el peso *medio* de la población de las RN hembras.

Como un segundo ejemplo consideremos la frase: «*El fumar incrementa el riesgo de hacer cáncer de pulmón*». Es cierta, aunque haya muchos casos particulares de fumadores empedernidos que llegan a viejos sin cáncer y otros casos de no fumadores que hacen cáncer, pues esa frase no alude a casos particulares, sino a que la *proporción* de cánceres es mayor en la *población de fumadores* que en la *población de no fumadores*.

Como tercer ejemplo, pensemos en la frase «*El fármaco “A” es hipotensor*». Esta afirmación es compatible con que haya algunos individuos hipertensos a los que ese fármaco no les baje la tensión arterial (TA), pues lo que dice es que la *media* de la TA de la *población* de hipertensos baja tras tomar ese fármaco.

Del mismo modo, se pueden seguir analizando otros muchos ejemplos, y en la mayoría de ellos se encontrará que:

Las afirmaciones del mundo de la biología consisten en afirmar algo acerca de las *medias o proporciones* de ciertas variables en determinadas *poblaciones*.

LA INFERENCIA ESTADÍSTICA INTENTA ELABORAR CONCLUSIONES RAZONABLES ACERCA DE LAS MEDIAS POBLACIONALES A PARTIR DE LAS MEDIAS MUESTRALES

En la mayoría de los casos, para poder hacer afirmaciones biológicas de validez general el investigador necesitaría conocer las *medias poblacionales* de ciertas variables, pero mediante experimentos con muestras solamente conoce las *medias de las muestras* estudiadas, y a partir de esos valores muestrales nunca puede conocer con exactitud los poblacionales.

Recordemos que una *población* es el total de individuos con ciertas características, y una *muestra* es una parte de esa población que nosotros observamos en una investigación concreta.

He aquí algunos ejemplos de poblaciones y de muestras:

1. La población de *varones adultos sanos españoles*: 16 millones.
 - Para intentar conocer cómo son los hábitos de higiene bucal en la población de varones adultos sanos españoles, hacemos una encuesta y exploración bucal en una *muestra de 1.234* varones adultos elegidos al azar entre toda la población.
2. La población de *enfermos de cierto tipo de úlcera de estómago*: 4 millones, en todo el mundo.
 - Para intentar saber cómo es la mucosa gástrica en este tipo de enfermos, analizamos las biopsias extraídas en una *muestra de 327* pacientes elegidos al azar entre toda la población que padece ese problema.
3. Veamos ahora un ejemplo en el que no es obvio cuál es la población.
 - Para intentar conocer si con tres miligramos de insulina cada día se puede compensar la ausencia de páncreas en la rata, se les extirpa el páncreas y se les administra esa dosis diaria de insulina a una *muestra de 20* ratas. En cada una de esas ratas se compara la glucemia basal antes y después del «tratamiento».
 - ¿Cuál es en este caso la población acerca de la cual queremos conocer? Esa población no existe físicamente, pues la forman el conjunto de todas las ratas que hubieran sido sometidas a ese mismo tratamiento, es decir, *extirpación de páncreas y administración de tres miligramos de insulina cada día*.

La Inferencia Estadística es el proceso de elaborar conclusiones razonables acerca de los valores poblacionales a partir del conocimiento de los datos muestrales. Y aunque es obvio que el conocimiento de la media de una *muestra* no permite, en general, saber cuánta es esa cantidad en la población correspondiente, a partir de la media muestral se pueden hacer dos cosas respecto al valor de la media poblacional:

- Calcular un *intervalo de confianza*, dentro del cual muy probablemente se encontrará el valor de la media de la población.

- Deducir, mediante un *test de significación*, que la media poblacional *no es* cierto valor, lo que puede ser de mucho interés para el conocimiento del tema investigado.

Ejemplo 1.º

Si en una *muestra de 2.000* ciudadanos encontramos que el 20% prefieren al Partido Liberal, todos entendemos que el correspondiente porcentaje en la *población*, el que aparece en las elecciones generales, no tiene por qué ser exactamente 20%. Aunque si la muestra es grande y está tomada aleatoriamente, el porcentaje (%) muestral será una estimación razonablemente buena del % de la población. El encontrar 20% de votantes del Partido Liberal en la muestra nos permite asumir que el % de votantes en la población será un valor relativamente próximo a 20% y que probablemente estará comprendido entre 10% y 30%. Y además es evidente que ese porcentaje poblacional no será, por ejemplo, 90%, ni 80,3%.

Ejemplo 2.º

Si la media de la concentración de L-arginina en tejido muscular es 230,5 unidades en una *muestra de 80* sujetos, la media poblacional, es decir, la media para todos los individuos de esa especie, probablemente estará dentro del intervalo 200-260 y, parece claro que *no será*, por ejemplo, 19 ni 865, ni otra cantidad muy alejada del valor 230,5 encontrado en la muestra.

TESTS DE SIGNIFICACIÓN E INTERVALOS DE CONFIANZA

Veamos sobre un nuevo ejemplo que la Inferencia Estadística es el intento de conocer acerca de los *parámetros* (medias o proporciones de las *poblaciones*) a partir de los *estadísticos* (medias o proporciones de las *muestras* estudiadas) y que se hace exactamente el mismo tipo de razonamiento que en la inferencia de la vida común.

En cada uno de estos países: España, Francia e Inglaterra, en 1.970 eran alérgicos a la aspirina (AA) el 8% de la población: $\Pi_{1970} = 0,08$. Se

sospecha que actualmente la proporción poblacional de AA quizá sea mayor de 8% en alguno de esos países. Para ver si actualmente ha aumentado esa proporción, y en cuánto lo ha hecho, tomaremos una muestra de $N = 1.000$ en cada país y veremos qué proporción de las personas de esa muestra son AA, y a la vista de esa información haremos un TS y calcularemos el IC.

Para hacer los TS planteamos, para cada país, la siguiente hipótesis nula, H_0 : *El porcentaje de AA no ha variado, es decir, $\Pi_{\text{POBLACIONAL ACTUAL}} = 0,08$* . Y la aceptaremos como posible o la rechazaremos según que el % de AA en la muestra sea compatible o no con ese valor poblacional.

El intervalo de confianza (IC) para el porcentaje actual de AA en esa población nos dice que tenemos cierta confianza (típicamente 95% o 99%, pero puede calcularse con cualquier otro nivel de confianza) en que dicho % poblacional esté dentro de ese intervalo.

1. España: en la muestra se encuentran 850 alérgicos: $P_{\text{muestral}} = 0,85$.

TS: ¿qué conclusión razonable tomaría ante este resultado?

- a) La Π_{Actual} *no es* 0,08 (rechazo H_0).
- b) La Π_{Actual} *puede ser* 0,08 (acepto H_0 como posible).
- c) La Π_{Actual} *es* 0,08 (afirmo que H_0 es cierta).

IC: en la muestra 85% $\rightarrow$ IC_{95%} para el % poblacional = 82,8% y 87,2%. Tenemos confianza de 95% en que el % poblacional esté entre 82,8% y 87,2%.

2. Francia: en la muestra se encuentran 90 alérgicos: $P_{\text{muestral}} = 0,09$.

TS: ¿qué conclusión razonable tomaría ante este resultado?

- a) La Π_{Actual} *no es* 0,08 (rechazo H_0).
- b) La Π_{Actual} *puede ser* 0,08 (acepto H_0 como posible).
- c) La Π_{Actual} *es* 0,08 (afirmo que H_0 es cierta).

IC: en la muestra 9% $\rightarrow$ IC_{95%} para el % poblacional = 7,3% y 10,9%. Tenemos confianza de 95% en que el % poblacional esté entre 7,3% y 10,9%.

3. Inglaterra: en la muestra se encuentran 80 alérgicos: $P_{\text{muestral}} = 0,08$.

TS: ¿qué conclusión razonable tomaría ante este resultado?

- a) La Π_{Actual} *no es* 0,08 (rechazo H_0).
- b) La Π_{Actual} *puede ser* 0,08 (acepto H_0 como posible).
- c) La Π_{Actual} *es* 0,08 (afirmo que H_0 es cierta).

IC: en la muestra 8% $\rightarrow$ IC_{95%} para el % poblacional = 6,4% y 9,8%. Tenemos confianza de 95% en que el % poblacional esté entre 6,4% y 9,8%¹.

El investigador decide la confianza que quiere tener en que el parámetro esté dentro del intervalo, pero cuanto más confianza quiere tener, más ancho quedará el intervalo. Lo más frecuente es dar el IC con 95% de confianza, pero aquí damos también los IC para el 99% y el 99,9% para el caso de Inglaterra.

$\rightarrow$ IC_{99%} para el % poblacional $\rightarrow$ 5,9% y 10,5%. Tenemos confianza 99% en que el % poblacional está entre 5,9% y 10,5%.

$\rightarrow$ IC_{99,9%} para el % poblacional $\rightarrow$ 4,8 % y 11,7%. Tenemos confianza 99,9% en que el % poblacional está entre 4,8% y 11,7%.

LA HIPÓTESIS DE TRABAJO, LA HIPÓTESIS NULA Y EL VALOR ESPERADO BAJO LA HIPÓTESIS NULA

Toda investigación surge a partir de una suposición que solemos llamar «hipótesis de trabajo». Por ejemplo, «Actualmente en España hay mayor proporción de alérgicos a la aspirina (AA) que en 1970». Aunque a primera vista no suele parecerlo, en la mayoría de los casos la hipótesis de trabajo alude a *medias poblacionales*, pero no proponiendo cuánto valen, sino cuánto *no valen*. Dice que la media de cierta variable en una *población* no es cierto número, sino una cantidad *mayor o menor* que cierto número.

Hipótesis de trabajo e hipótesis nula, H_0

En el ejemplo del apartado anterior la hipótesis de trabajo dice que la proporción poblacional actual de AA es «*mayor de 0,08*» y no especifica cuánto es exactamente, porque no hay razón para proponer una cifra con-

¹ «Soluciones correctas: a) España, b) Francia e Inglaterra».

creta. Lo relevante de esa hipótesis es el hecho de que el % de AA haya *aumentado* respecto al valor de 8% que tenía en 1970, y ello será cierto tanto si actualmente es, por ejemplo, 15%, lo que supondría un aumento moderado, como si es 49%, lo que supondría un aumento muy acusado.

Conocer el % observado en una muestra actual no nos permite saber cuánto vale exactamente el % en la población actual pero, aún sin saber cuánto vale exactamente el % poblacional, nuestra investigación será útil si de ella podemos deducir que *el porcentaje poblacional actual es mayor de 8%*, es decir, que ha aumentado.

Llamamos «hipótesis nula» y la simbolizamos por « H_0 » a la que establece que realmente no existe el efecto supuesto en la hipótesis de trabajo o, lo que es lo mismo, que en la población ese efecto es «nulo». En nuestro ejemplo la H_0 establece que el % de alérgicos es en la población actual 8%, es decir, que el aumento de AA es nulo. Resumiendo, en este ejemplo:

1. La *hipótesis de trabajo* dice que actualmente la proporción poblacional de AA es *mayor de 0,08*.
2. La *hipótesis nula*, H_0 , dice que el porcentaje de AA no ha variado, es decir, $\Pi_{\text{POBLACIONAL ACTUAL}} = 0,08$.

Igual que ocurría con los TS en la vida común, aceptaremos que la H_0 puede ser cierta si lo observado en la muestra es compatible con lo que ella propone. Diremos que una muestra es compatible con una hipótesis nula si de una población en la que se cumple la H_0 es fácil obtener una muestra del tipo de la observada en nuestro estudio. Y rechazaremos la H_0 si lo encontrado es incompatible o muy difícilmente compatible con la H_0 , es decir, si es muy difícil que de una población en que se cumple la H_0 salga una muestra del tipo de la que hemos obtenido en nuestra investigación.

Valor esperado y valor observado

Para evaluar si nuestra muestra es poco compatible con la H_0 , es decir, si es muy difícil que de una población en que se cumple la H_0 salga una muestra de ese tipo, comparamos el *valor esperado* bajo la H_0 con el *valor observado* en la muestra.

En nuestro ejemplo, la H_0 propone que son AA el 8% de la población y esperamos que en la muestra los AA sean aproximadamente el 8% y a este valor lo llamamos «porcentaje esperado». En una muestra de 1.000

personas el 8% es 80 y a esta cantidad la llamamos «valor esperado» de AA. El valor esperado es, pues, la cantidad de individuos de cierto tipo que tiene que haber en la muestra para que ella refleje exactamente lo que ocurre en una población en que se cumple la H_0 .

En la práctica, la mayoría de las muestras tomadas de una población en que se cumple la H_0 no van a contener exactamente dicho valor, pero sí valores próximos a él. En nuestro ejemplo, si tenemos una población con 8% de AA y tomamos muchas muestras al azar de 1.000 individuos cada una, el número de AA en ellas será exactamente 80 en algunas y valores próximos a esa cantidad en otras muchas. Y en pocas muestras tomadas de esa población aparecerá por azar una cantidad de AA muy alejada de 80, bien sea por encima o por debajo de ella.

Al número de AA encontrado en la muestra lo llamamos «*valor observado*» y la sospecha de que en la población no se cumple la H_0 es mayor cuanto mayor es la distancia entre el valor observado y el valor esperado.

Si en la muestra encontramos un valor observado muy alejado del esperado tendemos a pensar que la H_0 no se cumple en esa población, mientras que si obtenemos una muestra con valor observado próximo al esperado aceptamos que la H_0 *puede* ser cierta en la población.

Así, al encontrar en la muestra de 1.000 españoles 850 AA (es decir, 85% de AA) nos llevó a pensar que en la población de españoles *no hay* 8% de AA ya que el valor observado (85%) está muy alejado del esperado (8%). Y al encontrar en la muestra de 1.000 franceses 90 AA nos llevó a pensar que en la población de franceses *puede haber* 8%, pues 9% de AA observado se aleja muy poco del valor 8% esperado. Pero también es posible que el % poblacional sea otra cantidad no muy alejada del 9%, como los IC ponen de manifiesto.

EJEMPLOS DE HIPÓTESIS DE TRABAJO, HIPÓTESIS NULA Y VALOR ESPERADO BAJO LA HIPÓTESIS NULA

1.^{er} Ejemplo de hipótesis de trabajo, H_0 y valor esperado

Se sabe que son zurdos el 30% de los alemanes adultos. Sospechamos que el % de niños que nacen con esa característica es mayor del 30%, pero en parte de ellos es disfrazada por la educación. Puesto que en

los 10 últimos años ya no se actúa contra esa tendencia, si nuestra sospecha es cierta, el % de niños menores de 10 años zurdos será mayor del 30%. Para investigar el tema estudiamos a una muestra de $N = 20$ niños alemanes menores de 10 años, tomados al azar.

- *La hipótesis de trabajo* es que en la población actual de menores de 10 años el % de zurdos es mayor de 30.
- *La H_0* es que en esa población menor de 10 años son zurdos el 30% (es nula la diferencia entre esa generación y las anteriores).
- *El valor esperado bajo la H_0* es 6, porque esa cantidad es el 30% de la muestra de 20.

Criterios del TS: si el número de zurdos observados en la muestra es próximo a 6 pensaremos que ese dato es compatible con la H_0 , y si es muy superior a 6 pensaremos que la H_0 no es cierta, es decir, que el % de zurdos en la población de RN es mayor de 30.

2.^{do} Ejemplo de hipótesis de trabajo, H_0 y valor esperado

Con nuevas técnicas de densitometría ósea se midió la concentración de calcio (CC) en cuello humeral en todas las mujeres de más de 50 años, lo que permite conocer que la media poblacional es 300 (en unidades adecuadas) y desviación estándar 80. (Podrá seguir lo esencial de este razonamiento aunque no recuerde lo que es la desviación estándar).

Se sospecha que la halterofilia (HLT) incrementa la CC. El único modo de saberlo con certeza sería medir la CC a todas las mujeres que hacen HLT y ver si la media de esa población es o no mayor de 300. Pero no siendo eso posible medimos la CC a una muestra de $N=16$ levantadoras de peso. (Se descarta que este deporte pueda disminuir la CC).

Hipótesis de trabajo: la práctica de la HLT incrementa la CC en mujeres mayores de 50 años, es decir, la media de la CC en la población de mujeres que practican HLT es *mayor* que la media en la población de mujeres que no lo practican.

Hipótesis nula: el efecto de ese deporte sobre la CC es nulo, es decir, la media de la CC es *igual* en la población de las mujeres que practican HLT que en la población de mujeres que no lo practican.

Valor esperado bajo la H_0 : si la H_0 es cierta, en la muestra de 16 mujeres que practican este deporte esperamos que la media sea un valor próximo a 300.

Criterios del TS: si la media de CC observada en la muestra de 16 deportistas es próxima a 300, pensaremos que la H_0 puede ser cierta, el dato no constituye evidencia contra ella. Y si la media muestral es muy superior a 300, lo consideraremos evidencia contra la H_0 y a favor de que la media poblacional en la población de las que practican HLT es mayor de 300, es decir, ese deporte incrementa la calcificación ósea.

3.º Ejemplo de hipótesis de trabajo, H_0 y valor esperado

Al departamento de Psiquiatría de la Universidad de Duke llega el señor Zaj asegurando tener poderes telepáticos, de modo que puede adivinar el palo de las cartas de la baraja. Para investigar el tema se le invita a que diga el palo de cada una de las 100 cartas que se extraen al azar de las barajas españolas.

Hipótesis de trabajo: Zaj adivina el palo de las cartas más allá de las coincidencias esperadas por azar.

Hipótesis nula: Zaj no tiene esa capacidad y adivinará algunas veces el palo de la carta por simple azar.

Valor esperado bajo la H_0 : si no tuviera poderes especiales, como hay 4 palos cada vez que dice el de una carta tiene un cuarto de probabilidad de acertar por azar, de modo que si lo hace con 100 cartas esperamos que acierte aproximadamente en la cuarta parte de ellas, es decir, el valor esperado es 25.

Criterios del TS: si Zaj acierta, por ejemplo, en 28 cartas, consideramos que ese valor observado no se aleja mucho del esperado, 25, y hay acuerdo en decir que esa cifra es compatible con la H_0 , es decir, no constituye argumento fuerte a favor de que Zaj tiene esa capacidad. Si, por el contrario, Zaj acertara el palo de 90 cartas pensaríamos que ese valor observado se aleja tanto de 25 que es realmente difícil que haya aparecido por azar y consideraríamos ese resultado como indicativo de que Zaj tiene poderes telepáticos, es decir, lo consideramos una fuerte evidencia contra la H_0 .

RESULTADOS INTERMEDIOS: EL VALOR P DEL TEST

Hasta aquí hemos visto situaciones muy claras en las que todos los investigadores coinciden en las conclusiones que cabe hacer. En todos los ejemplos considerados hasta ahora el valor observado en cada estudio era o muy lejano al valor esperado (fuerte evidencia en contra de la H_0) o muy próximo al valor esperado (no evidencia contra la H_0).

Pero en muchos casos aparecen valores observados ni muy próximos ni muy alejados del valor esperado y no hay consenso en considerar si ese dato es o no evidencia seria contra la H_0 . Es en estos casos cuando se calcula el valor P del test, que puede ayudar a decantarnos por rechazar la H_0 o por aceptarla como posible.

Retomemos el ejemplo del apartado 3 en el que se decía que en 1.970 la proporción de personas con alergia a la aspirina (AA) era 0,08 en varios países: $\Pi_{1970} = 0,08$. Para ver si en alguno de ellos ha aumentado esa proporción tomamos una muestra de $N = 1.000$ en cada país, miramos la proporción muestral de AA e intentaremos elaborar conclusiones razonables acerca de si actualmente la proporción *poblacional* es o no mayor de 0,08.

En Francia se encontró $P_{\text{muestral}} = 0,09$ (en la muestra de 1.000 franceses, el 9% son AA), y en Inglaterra $P_{\text{muestral}} = 0,08$, lo que nos llevó a concluir que en ambos países la Π_{ACTUAL} podía seguir siendo 0,08, es decir, el dato muestral no es evidencia contra la H_0 . En España se encontró $P_{\text{muestral}} = 0,85$, lo que nos llevó a concluir que la Π_{ACTUAL} no es 0,08 sino mayor, porque el dato muestral es fuerte evidencia contra la H_0 .

Considere un cuarto país donde la muestra actual de $N = 1.000$ nos da $P_{\text{muestral}} = 0,17$ y un quinto país donde la muestra actual nos da $P_{\text{muestral}} = 0,13$. En estos casos el investigador no tiene claro hasta qué punto el resultado constituye o no evidencia fuerte contra la H_0 .

Es para ayudarnos en estas situaciones en las que calculamos el valor P del test. En el próximo capítulo veremos lo que indica ese valor P y por qué no siempre puede sacarnos de las dudas que nos llevaron a calcularlo. En unos casos nos obligará a decantarnos claramente contra la H_0 , en otros nos dirá que debemos aceptarla como posible, pero en otros casos no nos permitirá salir de la incertidumbre.

Ser consciente de estas limitaciones del valor P del test, saber que no siempre nos permite pronunciarnos a favor o en contra de una hipótesis,

es imprescindible para no caer en graves errores de concepto y de aplicación práctica.

COMPRUEBE SU NIVEL DE CONOCIMIENTOS: ENCUESTA DE AUTOEVALUACIÓN

En el Apéndice 2 encontrará una encuesta de autoevaluación para este capítulo, que le ayudará a evaluar en qué medida tiene claras sus ideas en este tema.

Interpretación del valor P de los tests de significación

Veíamos en el capítulo anterior que la esencia del razonamiento de los tests de significación (TS) consiste en rechazar una hipótesis cuando el resultado encontrado en el experimento es incompatible con ella. Valores observados muy lejanos al esperado bajo la H_0 constituyen fuerte evidencia en contra de ella, mientras que valores observados muy próximos al esperado no constituyen evidencia contra la H_0 , ni tampoco a su favor.

Pero en muchos casos el valor observado no es muy próximo pero tampoco muy alejado al esperado y no hay consenso en considerar si es o no evidencia seria contra la H_0 . Es en estos casos cuando se calcula el valor P del test, que puede ayudar a decantarnos por rechazar la H_0 o por aceptarla como posible.

En este capítulo explicamos lo que indica el valor P y por qué valores muy pequeños de esa probabilidad sugieren claramente que no es cierta la H_0 , mientras que valores no muy pequeños (medianos o grandes) no sugieren nada en particular. Veremos que no hay un valor concreto que separe los valores P «muy pequeños» de los «no muy pequeños», es decir, no hay una cifra frontera que separe los valores de P que llevan a rechazar la hipótesis de los que llevan a aceptarla como posible. Se trata de un proceso gradual en el que no hay un punto de corte.

También en este aspecto la lógica de los TS incorpora un *razonamiento típico de muchas situaciones de la vida común* en las que una magnitud puede tomar valores cualesquiera dentro de un rango y no

hay un valor frontera que marque una diferencia nítida entre dos zonas.

En primer lugar explicaremos cómo se interpretan esos valores que llamamos «Probabilidad» en el contexto de la investigación.

¿QUÉ INDICA LA PROBABILIDAD?

En Medicina, como en todas las ciencias aplicadas, «probabilidad» es sinónimo de frecuencia relativa o «tanto por uno», que también se puede expresar como «tanto por ciento» o porcentaje. No es, pues, una cantidad cuya explicación requiera conocimientos matemáticos, sino algo tan sencillo como decir, por ejemplo, que de cien individuos 42 están enfermos. Veamos algunos ejemplos clarificadores.

- 1.º ¿Qué quiere decir la expresión: «En los varones franceses de 60 años e hipertensos la probabilidad de hacer infarto de miocardio en los 12 meses siguientes es $P = 0,03$ »? Si multiplicamos ese valor por 100 nos da 3, es decir, «3 por ciento» y quiere decir que, en promedio, 3 de cada 100 personas con esas características (varones franceses hipertensos de 60 años) hacen infarto de miocardio en los 12 meses siguientes.
- 2.º ¿Qué quiere decir la expresión: «En cierta técnica quirúrgica la probabilidad de éxito es $P = 0,68$ »? Simplemente que de cada 100 veces que se ejecuta, hay éxito en 68.
- 3.º ¿Qué quiere decir la expresión: «Si una familia tiene tres hijos, la probabilidad de que alguno de ellos tenga problemas de drogadicción es $P = 0,09$ »? Simplemente que de cada 100 familias con tres hijos, en 9 de ellas alguno de los vástagos tiene ese tipo de problemas.
- 4.º ¿Qué quiere decir la expresión: «Si en una población hay 20% de enfermos y se toma una muestra aleatoria de $N = 5$, la probabilidad de que los cinco individuos sean enfermos es $P = 0,00032$ »? Que si se toman muchas muestras aleatorias de ese tamaño, en 32 de cada 100.000 de esas muestras son enfermos los cinco individuos.

Estas ideas tan sencillas deben ser recordadas muy claramente porque el valor P del test es una probabilidad y como tal puede expresarse

como porcentaje (o tanto por mil, o tanto por diez mil...) y para entender de modo claro y sencillo lo que indica hay que identificar quiénes son los 100 y qué les ocurre a cierta parte de ellos.

EL VALOR P DEL TEST. EJEMPLO CON RESULTADO EXTREMO: 1.º

Sabemos que en 1970 tenía la boca libre de caries el 20% de los escolares: $\Pi_{1970} = 0,2$. Tras una campaña educativa esperamos que el porcentaje (%) de niños con boca sana (BS) haya aumentado. En un primer estudio piloto se explora a $N = 5$ niños y se encuentra que los 5 tienen BS.

Hipótesis nula planteada, H_0 : en la población actual tienen BS el 20%, es decir, la campaña educativa no fue efectiva, tuvo eficacia «nula».

El hecho observado: en la muestra de $N = 5$ tienen BS el 100%.

Si el hecho observado es «difícilmente compatible» con la hipótesis, rechazaremos la H_0 , y pensaremos que actualmente el % poblacional de niños con BS es mayor de 20%, es decir, que la campaña educativa ha sido efectiva. Si el hecho es compatible con la H_0 , diremos que el % actual de niños con BS puede ser 20%, es decir, puede que la campaña educativa no haya sido efectiva, el hecho observado no es evidencia fuerte a favor de que la campaña ha incrementado el % de niños con BS. ¿Que salgan los 5 niños con BS es compatible con que solo el 20% de los niños tengan BS?

¡¡Recurramos a la práctica!!

Creamos una población donde realmente el 20% de los individuos tienen cierta característica. Sacamos muchas muestras al azar de $N = 5$ y contamos cuántas tienen los 5 individuos con la característica. Para ello se metieron en un bombo de lotería bolas blancas (20%) y negras (80%) y se sacaron 10 millones de muestras de 5 bolas cada una. Estas son las primeras 120 muestras:

	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0			
1	N	N	N	N	B	N	N	N	B	N	N	N	N	N	B	N	N	N	B	N	N	N	N	N	3	4	5	6	N	N	N	B	N
	N	N	N	N	N	N	N	N	B	B	N	B	N	N	B	N	N	B	N	B	N	N	N	N	N	N	N	N	N	N	B	B	
	N	B	N	N	N	N	N	N	N	N	N	N	B	N	N	N	B	N	N	N	N	N	N	B	N	N	N	N	N	N	N	N	
	N	N	B	N	B	N	B	N	N	B	B	B	N	N	B	N	N	N	N	N	N	N	N	N	B	N	B	N	N	N	N	B	
	N	B	B	N	N	B	N	N	N	N	N	B	N	B	N	N	N	N	N	N	N	N	N	B	B	N	N	B	N	N	N	N	
2	N	N	B	N	B	N	N	N	B	N	N	N	N	N	B	N	N	N	B	N	N	N	N	N	N	N	B	N	N	N	B	N	
	B	N	N	N	N	N	N	N	B	B	N	B	N	N	B	N	N	B	N	B	N	B	N	N	N	N	N	B	N	N	B	B	
	N	B	N	N	N	N	N	N	N	N	N	N	N	N	N	N	B	B	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
	N	N	N	N	B	N	N	N	N	B	B	B	N	N	B	N	N	N	N	N	N	N	N	N	B	N	B	N	N	N	N	B	
	N	N	B	N	N	B	N	N	N	N	N	B	N	B	N	N	N	N	N	N	N	N	N	B	B	N	N	B	N	N	N	N	
3	N	N	N	N	B	N	N	N	B	N	N	N	N	N	B	N	N	N	B	N	N	N	N	N	N	B	N	N	N	B	N	N	
	N	N	N	N	N	N	N	N	B	B	N	B	N	N	B	N	N	B	N	B	N	B	N	N	N	N	N	N	N	N	B	B	
	N	B	N	N	N	N	N	B	N	N	N	N	N	N	N	N	B	N	N	N	N	N	N	B	N	N	N	N	B	N	N	N	
	N	N	B	N	B	N	N	N	N	B	B	B	N	N	B	N	N	N	N	N	N	N	N	N	B	N	B	N	N	N	N	B	
	N	B	B	N	N	B	N	N	N	N	N	B	N	B	N	N	N	N	N	N	N	N	N	B	B	N	N	B	N	N	N	N	
4	N	N	N	N	B	N	N	N	B	N	N	N	N	N	B	N	N	N	B	N	N	N	N	N	N	N	B	N	N	N	B	N	
	N	N	N	N	N	N	N	N	B	B	N	B	N	N	B	N	N	B	N	B	N	B	N	N	N	N	N	N	N	N	B	B	
	B	B	N	N	B	N	B	N	N	N	N	N	N	N	N	N	B	N	N	N	N	N	N	B	N	N	N	N	N	N	N	N	
	N	N	B	N	N	N	N	N	N	B	B	B	N	N	B	N	N	N	N	N	N	N	N	N	B	N	B	N	N	B	N	B	
	N	B	B	N	N	B	N	N	N	N	N	B	N	B	N	N	N	N	N	N	N	N	N	N	B	B	N	N	B	N	N	N	

Vemos que entre esas 120 muestras no hay ninguna con las cinco bolas blancas ni tampoco con 4.

Sacando 10 millones de muestras se encuentran estas proporciones:

0 Blancas $\rightarrow$ 0,328 1 Blancas $\rightarrow$ 0,410 2 Blancas $\rightarrow$ 0,205
 3 Blancas $\rightarrow$ 0,051 4 Blancas $\rightarrow$ 0,006 5 Blancas $\rightarrow$ 0,0003

Es decir, solo 3 de cada 10.000 muestras tienen las 5 bolas blancas. El valor P del test es $P = 0,0003$.

Si la campaña educativa no consiguió aumentar el % de niños con BS, y por tanto actualmente sigue siendo 20%, al sacar 10.000 muestras de $N = 5$ niños, solo en 3 de ellas aparecen los 5 niños con BS.

Realmente no sabemos si el % actual de niños con BS es 20% o es mayor, pero vemos que si fuera 20% sería muy difícil que en una muestra de 5 todos tuvieran BS, y como en nuestra muestra ha sido así, lo razonable es pensar que el % actual poblacional es mayor de 20%, es decir, que la campaña ha sido efectiva.

Entre estas 30 no se encuentra ninguna con 4, ni tampoco con 3 blancas. Al hacer recuento sobre un millón de muestras se obtienen estos porcentajes:

	Con 0 Blanca	Con 1 Blanca	Con 2 Blancas	Con 3 Blancas	Con 4 Blancas
n_i	654.100	291.600	48.600	3.600	100
$f = n_i/N$	0,6541	0,2916	0,0486	0,0036	0,0001

Vemos que si en una población hay 10% de bolas blancas y se sacan muchas muestras aleatorias de $N = 4$ bolas, es muy improbable que aparezca una muestra con las cuatro blancas. Solo una cada 10.000 muestras tiene las 4 bolas blancas. Por la misma razón, si «A» cura realmente el 10% es muy improbable que en una muestra de $N = 4$ tratados con «A» se curen todos. Eso ocurre por azar en una muestra de cada 10.000. Es decir, si «A» cura realmente el 10% y trato muchas muestras de $N = 4$ con «A», solo en una de cada 10.000 muestras aparecerán curados los 4. Ese es el valor P del test.

Observando el % de curaciones en una muestra de 4 enfermos no podemos saber cuál es la proporción de curaciones que se obtendría si se tratara toda la población de enfermos, pero el hecho de que en la muestra aparezcan curados todos es difícilmente compatible con que «A» cure realmente el 10%. Por ello, el hecho de que en la única muestra tratada se hayan curado todos constituye un argumento bastante consistente a favor de que «A» cura realmente más del 10%, es decir, supone una mejora respecto a no tratar a estos pacientes.

Aunque no es el objeto de este capítulo, es obligado decir que al reportar este resultado lo correcto es dar, además del valor P del test, el intervalo de confianza (IC) para el % de curaciones que se obtendría si se tratara toda la población con «A». Un cálculo sencillo muestra que el $IC_{99\%}$ poblacional es 27% y 100%, es decir, tenemos confianza 99% en que el % de curaciones que se obtendría si se tratara toda la población de enfermos con «A» es un valor comprendido entre 27% y 100%.

EL VALOR P DEL TEST. EJEMPLO CON RESULTADO NO EXTREMO

En los dos ejemplos anteriores el resultado obtenido era el más extremo posible: en la muestra de 5 niños todos tenían la boca sana y en la

muestra de 4 enfermos tratados con «A» se curaron todos. A continuación veremos ejemplos donde el resultado no es tan extremo y veremos que en esos casos lo que se calcula es la probabilidad de obtener una muestra como la obtenida, *o con valor observado aún más alejado*.

Continuemos con un ejemplo en que intentamos saber si cada una de 5 monedas utilizadas para juegos de azar es o no equilibrada. (Llamamos «equilibrada» a una moneda si al lanzarla al aire la probabilidad de salir cara es 0,5, es decir, si al ser lanzada muchas veces la proporción de ellas en que aparece cara se aproxima progresivamente a 0,5).

Ante la sospecha de que alguna de ellas haya sido trucada se realiza una «investigación» consistente en lanzar cada moneda 200 veces y ver si el número de caras que aparecen es razonablemente próximo a 100, valor esperado si la moneda es equilibrada. Estos son los resultados:

Moneda	Número de lanzamientos	Número de caras esperadas	Número de caras observadas	% de caras observadas
A	200	100	180	90%
B	200	100	104	52%
C	200	100	100	50%
D	200	100	116	58%
E	200	100	134	67%

Al hacer TS (en el que planteamos la H_0 que dice que la moneda es equilibrada) para las tres primeras monedas todos los observadores estarían de acuerdo en la conclusión razonable que cabe sacar. Concluimos que la moneda «A» *no es* equilibrada, porque las 180 caras observadas se alejan mucho de las 100 esperadas si lo fuera, el dato es muy difícilmente compatible con la H_0 . Para la moneda «B» concluimos que *puede ser* equilibrada, porque las 104 caras observadas es una cantidad muy próxima a las 100 esperadas si la H_0 es cierta, es decir, el dato obtenido es compatible con la H_0 . El caso de la moneda «C» es equivalente al de la B y el hecho de que el valor observado en la muestra haya sido precisamente igual al esperado no nos lleva a decir que H_0 es cierta, sino que *puede serlo*.

Pero con la moneda «E» ya no hay consenso. Unos consideran que con una moneda equilibrada es muy difícil que salgan 134 caras y otros piensan que con una moneda equilibrada no es muy difícil que salgan 134 caras. ¿Qué se puede hacer en estos casos? ¿Cómo salir de esa duda?

Finalmente los investigadores llegaron a la conclusión más simple y razonable:

iii Recurrir a la práctica!!!

Es decir, se toma una moneda perfectamente equilibrada, se hacen muchas series de 200 lanzamientos cada una y se ve en qué proporción de esas series salen 134 o más caras.

¿Por qué se mira la proporción de «134 caras o más» en lugar de la proporción de series en las que han salido «exactamente 134 caras»? En el capítulo siguiente veremos esto con detalle, pero ya adelantamos aquí que este mismo tipo de razonamiento (mirar la proporción de casos tanto o más extremos que el encontrado) lo utilizamos continuamente en la vida cotidiana. A modo de adelanto vea el siguiente ejemplo.

A las 16:00 h están citadas con usted dos personas. Mr. A llega exactamente 15 segundos después de las 16:00 h y Mr. B llega exactamente 50 minutos y 12 segundos tarde. Todos consideramos el retraso de Mr. B inusual y punible, no porque muy pocas personas lleguen a las citas con ese retraso exactamente (50 minutos y 12 segundos), sino porque muy pocas personas se retrasan *tanto o más* que él. Por el contrario, Mr. A no es recriminado porque una elevada proporción de personas se retrasan *tanto o más* que él. Sin embargo, muy pocas personas llegan con un retraso exactamente igual que el de Mr. A, como también muy pocas llegaban con un retraso exactamente igual al de Mr. B.

Hay otras muchas situaciones de la vida común en las que, sin ser especialmente conscientes de ello, lo que evaluamos es la proporción de casos «iguales o más extremos que», no solamente «iguales que».

Igualmente los científicos optaron por evaluar lo difícil que sería obtener 134 caras por la proporción de series en las que se encuentran *134 o más caras*. A esa proporción, como sabemos, se la llama «probabilidad» de obtener 134 o más caras con una moneda equilibrada, y es el famoso valor P del test.

Si aparecen muy pocas series con esa característica (134 o más caras) diremos que ese resultado (134 caras) es difícilmente compatible con que la moneda sea equilibrada, y el hecho de haberlo obtenido con la moneda «E» nos sugeriría que no es equilibrada.

Por el contrario, si con la moneda equilibrada aparecen bastantes series con 134 o más caras, diremos que ese resultado (134 caras) es

fácilmente compatible con que la moneda sea equilibrada y el hecho de haberlo obtenido con la moneda «E» no habla en contra de que sea equilibrada. Diríamos que la distancia entre el valor esperado, 100, y el observado, 134, se alcanza fácilmente con una moneda equilibrada y el hecho de que con la «E» haya aparecido ese número de caras no lo consideramos un indicio fuerte de que sea sesgada.

Este criterio tan sencillo y razonable es la piedra angular de este ejemplo y lo será igualmente al hacer los tests de significación en todas las investigaciones biomédicas.

La comisión científica encargada de elaborar conclusiones razonables acerca de la moneda «E» ordenó que se cogiera una moneda perfectamente equilibrada y que se lanzara 200 veces y esto se repitiera *10 millones* de veces, es decir, se hicieran 10 millones de series, cada una de 200 tiradas, y se mirara cuántas de esas series tuvieron 134 o más caras. Se encontró que solamente 5 series dieron ese resultado, es decir, la proporción de series en que ocurre eso con una moneda equilibrada es: $P = 0,0000005$, este es el valor P del test.

«Por tanto —razonó la comisión científica— es muy difícil que al lanzar 200 veces una moneda buena, el valor observado sea 134 caras. Entonces lo razonable es pensar que la moneda «E» no es equilibrada sino que está sesgada a favor de las caras».

Es decir, el caso de la moneda «E» es equivalente al caso de la «A». Con la «A» no hizo falta hacer la prueba de tirar una moneda equilibrada muchas series de 200 veces cada una, porque había consenso general en que con una moneda equilibrada es realmente difícilísimo que salgan 180 caras.

El razonamiento usado en estos casos es paralelo al del caso del soldado Abel del capítulo anterior, en el sentido de que se rechaza la hipótesis planteada porque el dato encontrado es difícilmente compatible con ella.

Para la moneda «D» —en los 200 lanzamientos salieron 116 caras— se hace un recuento equivalente, es decir, una moneda perfectamente equilibrada se lanza 200 veces y esto se repite *10 millones* de veces y se observa en cuántas de esas series aparecieron 116 o más caras. Se encontró que 140.000 de las series tuvieron 116 o más caras, o lo que es igual, 14 por 1.000, es decir, la proporción de series en que ocurre eso con una moneda

buenas es $P = 0,014$. Este es el valor P del test. En este caso el cálculo de esa proporción no permite tomar postura definida, pues ese resultado de 116 caras no es muy fácil, pero tampoco muy difícil que aparezca con una moneda equilibrada. Diremos que el valor observado, 116, no es muy próximo al valor esperado, 100, pero tampoco es extremadamente alejado.

Veremos en un próximo capítulo que este tipo de incertidumbre ocurre muchas veces en las investigaciones reales y que el investigador serio debe asumirla como tal. Es inútil y contraproducente intentar salir de esa incertidumbre con falsas soluciones que no añaden información pero sí confusión.

Resumamos el resultado encontrado para cada moneda junto con la conclusión a la que llegamos inicialmente, el valor P del TS y la conclusión que nos sugiere dicho valor.

Recuerde que la H_0 planteada es que la moneda es equilibrada.

	% de caras observadas	Conclusión inicial	Valor P	Conclusión tras conocer P
A	90%	Rechazamos H_0	0,00.....01 con 30 ceros	Rechazamos H_0
B	52%	No rechazamos H_0	0,31	No rechazamos H_0
C	50%	No rechazamos H_0	0,50	No rechazamos H_0
D	58%	Dudoso	0,014	Dudoso
E	67%	Dudoso	0,00000005	Rechazamos H_0

Es decir, en las tres primeras monedas la interpretación del resultado es clara sin necesidad de calcular el valor P y cuando se calcula confirma lo que se había pensado previamente. En el caso de la moneda «E» la incertidumbre inicial se resuelve hacia el rechazo de la H_0 , pues es realmente difícil obtener una muestra de ese tipo por azar. En el caso de la moneda «D» la incertidumbre inicial no se resuelve al calcular el valor P del test, pues el resultado no constituye una evidencia fuerte contra esa hipótesis, aunque crea una duda razonable acerca de ella.

OTRAS HIPÓTESIS «NULAS»

Volvamos al caso de la moneda «E», que en 200 tiradas produjo **134** caras (67%) lo que nos llevó a rechazar tajantemente que fuera equilibra-

da, es decir, que a la larga produjera 50% de caras. Cabe plantearse otras hipótesis y actuar de modo equivalente.

- a) Pongamos como ejemplo que alguien sugiere que esa moneda produce a la larga 70% de caras. Ahora la «hipótesis nula» es precisamente esa, que a la larga esa moneda produce 70% de caras. Si una moneda que a la larga produjera cara en el 70% de la tiradas se lanza 200 veces esperamos que salgan unas **140** caras. Si se hacen muchas series de 200 tiradas cada una, en la mayoría de ellas saldrá un número de caras próximo a 140. ¿Abundan las muestras con 134 o menos caras? El valor P del test es *0,18*, es decir, en 18 de cada 100 series hay 134 caras o menos¹. Y en otras 18 de cada 100 series hay 146 caras o más (146 se aleja de 140 en 6 unidades, tanto como se aleja 134 pero en diferente sentido). Vemos que obtener 134 caras es compatible con la hipótesis nula que hemos planteado y por ello concluimos que esa hipótesis puede ser cierta, es decir, que la moneda puede que a la larga produzca 70% de caras.
- b) Pongamos ahora como ejemplo que alguien sugiere que esa moneda produce a la larga 80% de caras. Ahora la «hipótesis nula» es precisamente esa, que a la larga esa moneda produce 80% de caras. Si una moneda que a la larga produjera cara en el 80% de las tiradas se lanza 200 veces esperamos que salgan unas 160 caras. Si se hacen muchas series de 200 tiradas cada una, en la mayoría de ellas saldrá un número de caras próximo a 160. ¿Abundan las muestras con 134 o menos caras? El valor P del test es *0,000002*, es decir, solo en 2 de cada millón de series hay 134 caras o menos. Y en otras 2 hay 186 caras o más (186 se aleja de 160 en 26 unidades, tanto como se aleja 134 pero en diferente sentido). Vemos que obtener 134 caras es difícilmente compatible con la hipótesis nula que hemos planteado y por ello nos inclinamos a pensar que esa hipótesis no es cierta, es decir, que a la larga la moneda no produce 80% de caras, sino menos.

¹ Recuerde que se trata de ver la proporción de muestras que se alejan de lo que propone la H_0 tanto o más que la obtenida, en nuestro ejemplo, proporción de muestras con 134 caras o cualquier otra cantidad más alejada de 140 en ese mismo sentido.

La idea es que el % muestral, en este caso 67%, no nos permite conocer el % poblacional (el que se obtendría a la larga) pero podemos ver en qué medida ese resultado es incompatible con un valor que se proponga para la población.

EPÍLOGO

Ahora ya sabemos lo que indica el valor P del test y cómo nos ayuda a elaborar conclusiones. Sabemos que el valor P del test es la probabilidad de obtener muestras que se alejen de la esperada bajo la H_0 tanto o más de lo que se aleja la encontrada en nuestro estudio, si en la población es cierta la hipótesis nula. Dado el papel central que este concepto tiene en la inferencia y las dificultades que con él tienen muchos investigadores, insistamos en estos puntos clave:

- a) En el cálculo de ese valor P pueden intervenir fórmulas matemáticas más o menos complicadas, y para eso están los estadísticos y los ordenadores, pero *entender lo que ese valor indica* está al alcance de todo profesional de la investigación, aunque tenga muy olvidados sus conocimientos matemáticos. El objetivo de este informe es, precisamente, explicar la interpretación de ese valor, no enseñar a calcularlo.
- b) El valor P del test es una probabilidad y, como veíamos en un apartado anterior de este capítulo, debemos interpretarlo sencillamente como un porcentaje, entendiendo claramente de qué cien (o mil o diez mil...) entidades se trata y qué les ocurre a algunas de ellas.
- c) En muchos casos el valor P del test no permite tomar una postura clara sobre el tema que se está investigando, como ocurría en el ejemplo de la moneda «D». Por tanto NO es la panacea universal, sino una ayuda parcial que hay que valorar en su justa medida y sin magnificar su importancia.

Probabilidad de un valor particular *versus* probabilidad de cola

Ya conocemos el proceso lógico de los tests de significación (TS), tanto en la vida común como para elaborar conclusiones razonables en la investigación científica con ayuda del valor P del test. La idea básica es tan simple como que se rechaza una hipótesis cuando los resultados del experimento son incompatibles o muy difícilmente compatibles con ella. En Inferencia Estadística el grado de compatibilidad o incompatibilidad entre hipótesis y resultados se expresa por el valor P del test, que es la probabilidad de encontrar una muestra con valor observado tan alejado del esperado bajo la hipótesis nula como el de nuestro estudio *o más*, si la H_0 es cierta¹.

Es lógico preguntarse por qué no se evalúa, simplemente, la probabilidad de obtener por azar una muestra *como la obtenida* en nuestro estudio, por qué se calcula también la probabilidad de muestras con valor observado aún *más alejado*.

Mostraremos que ese mismo proceso mental —evaluar la incompatibilidad de un dato con una hipótesis calculando la frecuencia relativa de datos *tan extremos como ese o aún más extremos*— lo usamos espontáneamente *en la vida común*. De modo que también en este punto la Inferencia Estadística aplica el mismo proceso lógico que usamos en la calle. Lo veremos a través de ejemplos prácticos.

¹ Y esa probabilidad es la proporción de muestras de ese tipo que aparece si se sacan muchas muestras del mismo tamaño de una población en la que se cumple lo que dice la H_0 .

EN LA VIDA COMÚN EVALUAMOS PROPORCIÓN DE CASOS COMO EL OBSERVADO O AÚN MÁS EXTREMOS

Ejemplo 1.º Estaturas normales y estaturas inusuales

Si el señor «A» mide 168,456 cm (es decir, 168 cm, 4 milímetros, 5 décimas de milímetro y 6 centésimas de milímetro) y el señor «B» mide 218,987 cm, consideramos la estatura del primero muy «normal» y la del segundo «extraordinariamente alta». ¿Y por qué las consideramos «normal» y «extraordinaria» respectivamente?

La respuesta correcta parece que es: «porque hay mucha gente con la estatura del señor «A» y muy poca con la de «B»». Pero eso no es totalmente cierto. Porque al estar registradas esas estaturas con tanta precisión en realidad también hay muy pocas personas que midan exactamente 168,456 cm. Lo que sí es cierto es que hay muchas personas con estaturas *próximas a la de «A»* y muy pocas con estaturas *próximas a la de «B»*. También es cierto que hay muchas personas con estatura superior a la de «A» y muy pocas con estatura superior a la de «B». Un modo de expresar cuán inusual es una estatura entre los individuos de un colectivo es dar la proporción de individuos con estatura superior a esa si es una persona alta, o inferior a esa si es baja, es decir, con estatura más extrema que esa.

Ejemplo 2.º Demoras aceptables y demoras intolerables

A las 16:00 están citadas con usted dos personas. «A» llega exactamente 15 segundos después y «B» 50 minutos y 12 segundos tarde. Todos consideramos el retraso de «B» inusual y punible. ¿Por qué? No porque muy pocas personas lleguen a las citas con ese retraso exactamente (50 minutos y 12 segundos), sino porque muy pocas personas se retrasan tanto o más que él. Por el contrario, «A» no es recriminado ya que una elevada proporción de personas se retrasan tanto o más que él. Sin embargo muy pocas personas llegan con un retraso exactamente igual que el de «A», como también muy pocas llegaban con un retraso exactamente igual al de «B».

Ejemplo 3.º ¿Han retrasado la hora de salida del avión procedente de Macondo?

El avión procedente de Macondo tiene su hora de salida a la 1:00 y llega a Barajas en torno a las 8:00, ambas horas de Madrid.

Sospechamos que hoy salió con retraso, pero no podemos averiguar la hora de salida y tenemos que emitir opinión al respecto teniendo como única información la hora de llegada. Si nos dicen que llegó a las 08:06:34 (las 8 horas, 6 minutos y 34 segundos) hay consenso en que ese dato no es una evidencia fuerte contra la hipótesis que dice que salió a la 1:00 y, por tanto, no lleva a rechazarla. Por el contrario, si el avión llega a las 17:23:06 hay consenso en que ese dato es una fuerte evidencia contra la hipótesis que dice que salió a la 1:00, y por tanto nos lleva a rechazarla.

¿Por qué la llegada a las 17:23:06 es una fuerte evidencia contar la H_0 (rechazamos que haya despegado a la 1:00) y la llegada a las 08:06:34 no es una fuerte evidencia contar la H_0 (no rechazamos que haya despegado a la 1:00)?

La respuesta obvia es: «porque si sale a la 1:00 es fácil que llegue a las 08:06:34, pero es muy difícil que llegue a las 17:23:06». Sin embargo, la proporción de aviones que saliendo a la 1:00 llegan exactamente a las 08:06:34 es pequeñísima. Son muchos los que llegan *cerca de esa hora*, pero casi ninguno llega justamente con ese retraso de 6 minutos y 34 segundos. Y no nos parece un retraso especialmente sospechoso porque son muchos los aviones que habiendo salido a la 1:00 llegan con *más retraso* que ese.

También son muy pocos los aviones que habiendo salido a la 1:00 llegan exactamente a las 17:23:06, pero lo que nos lleva a rechazar que haya salido a la 1:00 si llega a las 17:23:06 es el hecho de que muy pocos aviones se retrasan tanto como eso (9 horas, 23 minutos y 6 segundos) o *más*.

EN LA INFERENCIA ESTADÍSTICA LO RAZONABLE ES EVALUAR PROPORCIÓN DE CASOS COMO EL ENCONTRADO O AÚN MÁS EXTREMOS

Veámoslo a través de un ejemplo. ¿Son mujeres el **50%** de los recién nacidos actualmente en ciertos países de Europa?

Se sospecha que actualmente nacen en cada uno de 6 países más mujeres que varones. Para aportar luz sobre este tema se estudia una muestra de $N = 10.000$ recién nacidos (RN) en cada país. Para cada uno de ellos se plantea la H_0 que dice que *ambos sexos son igual de frecuentes*, o lo que es lo mismo, que son mujeres el 50% de la población de RN.

Si esta H_0 es cierta la cantidad de mujeres esperada en la muestra de 10.000 nacimientos es 5.000, y si se tomaran muchas muestras de ese tamaño en la mayoría de ellas aparecería un número de mujeres próximo a 5.000. Por ello, si en una muestra encontramos, por ejemplo, 5.003 o 5.007 lo consideramos un resultado compatible con la H_0 y que no constituye evidencia contra ella. Por el contrario, si en la muestra de un país fueran mujeres 9.800, por ejemplo, habrá consenso en considerar que este resultado es prácticamente incompatible con la H_0 , de modo que nos obliga a rechazarla y concluir que el % de mujeres en la población de RN de ese país es superior al 50%.

¿Y que pensaríamos si encontramos 5.050 mujeres en la muestra de 10.000 RN? Si calculamos la probabilidad de encontrar 5.050 mujeres en una muestra tomada de una población donde realmente hay 50% de mujeres, se obtiene $P \approx 0,004^2$. Al ser una proporción pequeña podría parecer que ese hallazgo habla a favor de que en la población muestreada no hay 50% de mujeres, sino más. Pero hay que tener en cuenta que al ser una muestra bastante grande hay muchos resultados posibles y cada uno de ellos tiene una probabilidad bastante baja. Por ejemplo, el valor 5.020, que es mucho más próximo a la mitad de la muestra, tiene una probabilidad de $P = 0,007$ y el valor 5.000, justamente el valor teórico esperado en una población en que haya 50% de mujeres, tiene $P = 0,008^3$. Obviamente, si en la muestra aparecen precisamente 5.000 mujeres no lo consideramos un argumento contra la H_0 , sino que es el resultado más favorable en relación con esa H_0 , y sin embargo la probabilidad de obtener una muestra con ese número concreto de mujeres es solamente $P = 0,008$. Por tanto:

Una muestra puede ser totalmente compatible con una hipótesis y sin embargo tener una probabilidad muy baja de aparecer al muestrear una población en que se cumpla esa hipótesis.

Por ello, el hecho de que una muestra sea muy improbable no es argumento contra la H_0 . Entonces, ¿qué magnitud constituye evidencia contra la H_0 ? Ya vimos en situaciones equivalentes a estas en la vida

² Es decir, de cada 1.000 muestras (cada una con 10.000 RN) 4 tendrían 5.050 mujeres.

³ Es decir, al tomar muestras de 10.000 RN de una población donde son mujeres la mitad de ellos, de cada mil muestras solo 8 tendrán exactamente la mitad de mujeres.

común que lo que se tiene en cuenta es la proporción de muestras más extremas⁴ que la encontrada en nuestro estudio.

Pensamos que una muestra es compatible con una hipótesis si es grande la probabilidad obtener muestras como esa *o aún más extremas*.

En la siguiente tabla se dan los resultados obtenidos al extraer una muestra de 10.000 RN en cada uno de los seis países: A, B, C... F. Para cada país se da el número de mujeres en la muestra, la impresión subjetiva a partir de este resultado muestral, y las probabilidades de obtener, en una población en que se cumpla la H_0 :

- Una muestra como la encontrada (columna tercera).
- Una muestra como esa o aún más extrema (columna cuarta).

La última columna de la tabla contiene la impresión del investigador a la vista de los valores P obtenidos.

Vea detenidamente el resultado de cada país y observe en qué medida el conocimiento del valor P modifica nuestra opinión.

Recuerde que para cada país la H_0 dice que ambos sexos son igual de frecuentes en población.

País	Número de mujeres en la muestra de 10.000 RN	¿Es la muestra compatible o incompatible con la H_0 ? <i>Impresión subjetiva</i>	Proporción de muestras con ese número de mujeres si en la población son mujeres el 50%	Proporción de muestras con ese número de mujeres o más, si en la población son mujeres el 50%	¿Es la muestra compatible o incompatible con la H_0 ? <i>Opinión al ver el valor P</i>
A	5.000	Totalmente compatible	0,008	0,50	
B	5.010	Compatible	0,007	0,42	
C	5.050	???	0,004	0,15	Compatible
D	5.070	???	0,003	0,08	Compatible
E	5.250	???	0,00000006	0,0000003	Muy poco compatible
F	9.800	Claramente incompatible	10^{-301}	10^{-300}	Claramente incompatible

⁴ Por muestras «más extremas» entendemos, obviamente, las que tienen valores observados más alejados del valor esperado bajo la H_0 .

En el **país «A»** se encuentran precisamente 5.000 mujeres en la muestra, que es el valor esperado bajo la H_0 . Esta muestra, más que ninguna otra, es compatible con la H_0 , y sin embargo la probabilidad de obtener exactamente esa muestra es muy pequeña⁵ (solo 8 de cada 1.000 muestras tendrán exactamente 5.000 mujeres).

En el **país «B»** se encuentran 5.010 mujeres, lo que es totalmente compatible con la H_0 , pues 5.010 es un valor muy cercano a las 5.000 esperadas.

- Si de una población con 50% de mujeres tomamos muestras de 10.000 individuos, el porcentaje de ellas en que aparecen *exactamente* 5.010 mujeres es solo 0,007 (7 de cada 1.000). Esta baja probabilidad no constituye evidencia contra la H_0 (cualquier otra muestra tiene una probabilidad muy baja).
- En esa misma población la proporción de muestras con 5.010 o *más* mujeres es 0,42 (42 cada 100). Esta probabilidad indica la compatibilidad entre la muestra y la H_0 , es la P del test.

Decimos que la muestra con 5.010 mujeres es compatible con la H_0 porque es bastante probable obtener una muestra con 5.010 o *más* mujeres.

En el **país «C»** encontramos una muestra con 5.050 mujeres, y usando solo la intuición ya no hay unanimidad en pensar si es o no compatible con la H_0 . La probabilidad de que salgan justamente 5.050 mujeres es solo 0,004, pero la probabilidad de que salgan 5.050 o *más* es 0,15 y a la vista de esa probabilidad la consideraremos compatible con la H_0 .

En el **país «D»** aparece una muestra con 5.070 mujeres, no hay unanimidad en decir si es o no compatible con la H_0 . Tiene probabilidad de salir de solo 0,003, pero la probabilidad de que salgan 5.070 o *más* es 0,08 y a la vista de esa probabilidad la consideraremos compatible con la H_0 .

En el **país «E»** aparece una muestra con 5.250 mujeres. La probabilidad de salir es solo 0,00000003, y la probabilidad de que salgan 5.250 o *más* es 0,00000003, y a la vista de esa probabilidad la consideraremos prácticamente incompatible con la H_0 .

En el **país «F»** se encuentran 9.800 mujeres, lo que a todas luces es prácticamente imposible que ocurra en una muestra tomada de una

⁵ «Aunque lo relevante es el concepto y no el cálculo, esto es una binomial de $N = 10.000$ y $\Pi = 0,5$ ».

población con 50% de mujeres. En estos casos no es necesario calcular el valor P. El sentido común da la pauta adecuada.

Resumiendo, consideramos que una muestra es difícilmente compatible con una hipótesis si es pequeña la probabilidad de obtener por azar muestras como la obtenida en nuestro estudio *o aún más alejadas* de lo esperado bajo la hipótesis nula. Ese es el valor P del test, que nos dice qué proporción de muestras tienen valor observado tan alejado del esperado o aún más alejado, si se extraen muchas muestras de una población en la que se cumple la hipótesis nula. Ese mismo criterio se sigue en la vida común.

COMPRUEBE SU NIVEL DE CONOCIMIENTOS: ENCUESTA DE AUTOEVALUACIÓN

En el Apéndice 2 encontrará una encuesta de autoevaluación para este capítulo, que le ayudará a evaluar en qué medida tiene claras sus ideas en este tema.

Más ejemplos de interpretación del valor P del test

NOTA PREVIA: en los capítulos anteriores vimos lo que indica el valor P del test a través de varios ejemplos. Dado el papel central que el concepto y la interpretación de esa cantidad juegan al elaborar las conclusiones de los trabajos científicos, se incluye este capítulo con más ejemplos sobre el mismo tema. No añade nada nuevo. Será útil para quienes ven estos conceptos por vez primera y necesiten afianzarlos.

EL VALOR P DEL TEST. EJEMPLO CON VALORES EXTREMOS

En cada uno de estos países, España, Francia, Italia e Inglaterra, en 1980 la proporción de personas adictas al tabaco (AT) era $\Pi_{1980} = 0,5$. Para ver si actualmente ha aumentado esa proporción tomaremos una muestra en cada país y veremos qué proporción de las personas de esa muestra son AT. En todos ellos planteamos como hipótesis nula H_0 : $\Pi_{\text{ACTUAL}} = 0,5$, es decir, no ha variado la proporción de AT en la población. ¿Qué concluimos según el resultado de cada país.

1.º España: en una muestra de $N = 4$ españoles se encuentra que son AT los 4. ¿Cuál es la conclusión razonable?

a) Rechazo H_0

b) Acepto H_0 como posible

La P del test 0,06, es decir, si actualmente en España son AT el 50% y tomo muchas muestras de $N = 4$, en 6 de cada 100 aparece por azar que los 4 son AT. Lo que nos confirma la impresión de que este dato no es

una evidencia fuerte contra la H_0 . Y en ese mismo criterio nos coloca el cálculo del intervalo de confianza: $IC_{95\%}(\Pi_{\text{Actual}}) = (38\%, 100\%)$. Es decir, tenemos confianza de 95% en que el porcentaje de AT en la población actual es una cantidad comprendida entre 38% y 100%.

2.º Francia: en una muestra de $N = 100$ franceses se encuentra que son AT los 100. ¿Cuál es la conclusión razonable?

a) Rechazo H_0

b) Acepto H_0 como posible

La P del test 7×10^{-31} , es decir, mucho menos de 7 cada billón de billones. Si actualmente en Francia son AT el 50% y tomo muchas muestras de $N = 100$, en menos de 7 de cada billón de billones aparece por azar que los 100 son AT. Lo que nos confirma la impresión de que este dato es una evidencia prácticamente definitiva contra la H_0 . Y en ese mismo criterio nos coloca el cálculo del intervalo de confianza. $IC_{95\%}(\Pi_{\text{Actual}}) = (96\%, 100\%)$.

3.º Inglaterra: En una muestra de $N = 20$ ingleses se encuentra que son AT los 20. ¿Cuál es la conclusión razonable?

a) Rechazo H_0

b) Acepto H_0 como posible

En este caso puede ayudar conocer la P del test, que resulta ser 0,000001. Es decir, si actualmente en Inglaterra son AT el 50% y tomo muchas muestras de $N = 20$, en 1 de cada millón aparece por azar, que los 20 son AT.

Recuerde que cuando no está claro si el dato empírico es compatible con la hipótesis (porque no es muy fácil, pero tampoco muy difícil, que se dé ese dato si la hipótesis es cierta) lo más sensato es:

¡¡¡recurrir a la práctica!!!

Se trata de crear un sistema en el que se cumple la hipótesis que plantea el test, repetir el proceso de observación empírica muchas veces y ver con cuánta frecuencia ocurre lo que nos ocurrió en nuestro experimento. Si ese hecho ocurre con gran frecuencia decimos que es compatible con la H_0 , y si ocurre con escasa frecuencia decimos que es difícilmente compatible con la H_0 .

En el caso de Inglaterra (de 20 personas, todas son AT) crearemos una población donde realmente el 50% de las personas tienen cierta característica. Sacaremos millones de muestras de $N = 20$ y contaremos el número de ellas que tienen todos los individuos con esa característica.

Pero es equivalente a eso y mucho más sencillo usar una moneda equilibrada, hacer millones de series de $N = 20$ tiradas cada una y contar el número de series en las que sale cara en las 20 tiradas.

Los estadísticos lo han hecho por nosotros y han encontrado que ese hecho (cara en las 20 tiradas) ocurre en una serie cada millón.

$$\text{Frecuencia relativa} = 1 \text{ por } 1.000.000 = 0,000\ 001$$

Este es el valor P del test. Lo que nos lleva a pensar que este dato es una fuerte evidencia contra la H_0 . Y en ese mismo criterio nos coloca el cálculo del intervalo de confianza: $IC_{95\%}(\Pi_{\text{Actual}}) = (83\%, 100\%)$.

4.º Italia: En una muestra de $N = 12$ italianos se encuentra que son AT los 12. ¿Cuál es la conclusión razonable?

a) Rechazo H_0

b) Acepto H_0 como posible

También en este caso puede ayudar conocer la P del test, que resulta ser 0,0002, es decir, si actualmente son AT el 50% y tomo muchas muestras de $N = 12$, en 2 de cada 10.000 aparece por azar, que los 12 son AT. Por tanto el resultado constituye evidencia relativamente fuerte a favor de que actualmente los AT son más del 50%. Y en ese mismo criterio nos coloca el cálculo del intervalo de confianza: $IC_{95\%}(\Pi_{\text{Actual}}) = (74\%, 100\%)$.

Ciertamente, los estadísticos no han tenido que tirar la moneda millones de veces. Un razonamiento lógico muy sencillo nos dice con qué frecuencia aparece ese tipo de muestra bajo esa condición.

MÁS EJEMPLOS SOBRE EL VALOR P DEL TEST. RESULTADOS NO EXTREMOS

Veamos otros ejemplos en los que el resultado no es extremo, de modo que calculamos la probabilidad de obtener muestras con valor observado como el de nuestro experimento o *aún más extremo*.

En cada uno de estos países: España, Francia, Italia, Alemania y Rusia en 1980 la proporción de personas que consumían antidepresivos (CA)

era $\Pi_{1980} = 0,40$. Para ver si actualmente ha aumentado esa proporción tomaremos una muestra en cada país y veremos qué proporción de personas de esa muestra son CA. En todos ellos planteamos como hipótesis nula, $H_0: \Pi_{POBLACIONAL\ ACTUAL} = 0,40$, es decir, no ha variado la proporción de CA. Y veremos si el resultado muestral parece constituir un argumento contra ella o bien es compatible con ella, esto es, haremos un TS.

1.º España: se estudian 10 individuos ($N = 10$). Si $\Pi_{ACTUAL} = 0,40$, el valor esperado de CA es $E = 4$. En la muestra se encuentra que hay **6** CA, es decir, $\%_{MUESTRAL\ DE\ CA} = 60\%$.

TS $\rightarrow$ El valor observado, 6, está muy próximo al esperado, de modo que no parece que ese resultado sea evidencia seria contra la H_0 . Concluimos que el resultado es compatible con que en la población siga habiendo 40% de CA, es decir, que no haya aumentado ese porcentaje respecto al año 1980.

Si calculamos el valor de la P del test se obtiene $P = 0,17$ o 17%. ¿Qué quiere decir ese 17%? Si de una población en la que hay 40% de individuos con cierta característica (en nuestro caso, que son CA) sacamos muchas muestras de 10 individuos cada una, 17 de cada 100 muestras tienen 6 o más individuos con esa característica. Por ello, si en una muestra de $N = 10$ tomada de una población con % de CA desconocido hay 6 individuos que son CA pensamos que en esa población puede haber 40% de individuos CA.

2.º Francia: Muestra estudiada de $N = 100$ franceses. Si $\Pi_{ACTUAL} = 0,40$, el valor esperado es $E = 40$.

En la muestra se encuentra que hay **92** CA, es decir, $\%_{MUESTRAL} = 92\%$.

TS $\rightarrow$ El valor observado, 92, se aleja mucho del esperado, de modo que ese resultado constituye evidencia seria contra la H_0 . Concluimos que el resultado no es compatible con que en la población de franceses siga habiendo 40% de CA, es decir, pensamos que en la población de Francia los CA son más del 40%.

Si calculamos el valor de la P del test se obtiene $P = 0,00....003$ (con 24 ceros). Esto significa que si de una población en la que hay 40% de individuos con cierta característica sacamos muchas muestras de 100 individuos cada una, solo 3 de cada billón de billones de muestras tiene 92 o más individuos con esa característica. Es decir, si en la población hay 40% de CA es prácticamente imposible que en la muestra aparezcan

92%. Y como hemos obtenido 92% en la muestra, estamos prácticamente seguros de que no hay 40% en la población.

3.º Italia: muestra estudiada de $N = 60$ italianos. Si $\Pi_{\text{ACTUAL}} = 0,40$, el valor esperado es $E = 24$.

En la muestra se encuentra que hay **44** CA, es decir, $\%_{\text{MUESTRAL}} = 73,3\%$.

TS → El valor observado, 44, no está muy alejado del esperado, pero tampoco está muy cercano a él. Algún investigador puede pensar que el haber encontrado 44 CA debe ser interpretado como un fuerte argumento contra la H_0 , pero otros pueden considerar que ese resultado es compatible con la H_0 .

Es en estos casos cuando el cálculo del valor P puede ayudar decisivamente. Lo calcularemos construyendo una población de individuos en la que el 40% tiene cierta característica y tomando muchas muestras al azar de 60 individuos cada una y contando en cuantas de ellas aparecen *44 o más* individuos con esa característica. Se encuentra el valor $P = 0,0000001$, es decir, de cada diez millones de muestras solo una tienen 44 o más individuos con la característica. Siendo tan pequeña esa proporción, hay consenso general entre los investigadores en considerar el dato como una fuerte evidencia contra la H_0 , de modo que asumimos que en Italia son CA más del 40%.

4.º Alemania: la muestra estudiada es de $N = 50$. Si $\Pi_{\text{ACTUAL}} = 0,40$, el valor esperado es $E = 20$.

En la muestra se encuentra que hay **34** CA, es decir, $\%_{\text{MUESTRAL}} = 68\%$.

TS → El valor observado, 34, no está muy alejado del esperado, pero tampoco está muy cercano a él. Algún investigador puede pensar que el haber encontrado 34 CA debe ser interpretado como un fuerte argumento contra la H_0 , pero otros pueden considerar que ese resultado es compatible con ella.

Calcularemos el valor P construyendo una población de individuos en la que 40% tiene cierta característica, tomando muchas muestras al azar de 50 individuos cada una y contando en cuantas de ellas aparecen 34 o más individuos con esa característica. Se encuentra el valor $P = 0,00005$, es decir, de cada cien mil muestras solo 5 tienen 34 o más individuos con la característica. Este valor se considera una notable evidencia contra la H_0 .

5.º Rusia: la muestra estudiada es de $N = 50$. Si $\Pi_{\text{ACTUAL}} = 0,40$, el valor esperado es $E = 20$.

En la muestra se encuentra que hay **28 CA**, es decir, $\%_{\text{MUESTRAL}} = 56\%$.

TS → El valor observado, 28, no está muy alejado del esperado, pero tampoco está muy cercano a él. Calculamos el valor P construyendo una población de individuos en la que 40% tiene cierta característica, y tomando muchas muestras al azar de 50 individuos cada una y contando en cuántas de ellas aparecen 28 o más individuos con esa característica. Se encuentra el valor $P = 0,015$, es decir, de cada mil muestras 15 tienen 28 o más individuos con la característica. Este valor de P no constituye una fuerte evidencia contra la H_0 , pero sí da lugar a una justificada sospecha acerca de su veracidad.

¿Y cómo salir de la duda en este caso y otros similares? No hay modo de salir de la duda a partir de este resultado. La comunidad científica toma nota de este dato, que nos hace dudar de la veracidad de la H_0 , y espera que nuevos estudios sobre este tema confirmen (porque aparecen resultados en la misma dirección) o más bien desvanezcan esa sospecha.

Es imprescindible que el lector se dé cuenta de que en estos casos no se puede tomar postura y lo único que cabe es tener en cuenta el resultado para integrarlo con otros sobre el mismo tema. Y si no hubiera más información al respecto la incertidumbre se mantendría.

Test de significación comparando dos medias y dos proporciones

Para explicar la lógica de la Inferencia Estadística en la investigación médica hemos considerado hasta aquí siempre la situación más sencilla y típica en que a partir de la media o proporción de una muestra se calculan intervalos de confianza (IC) para el correspondiente valor poblacional y se hacen tests de significación (TS) para ver en qué medida el valor muestral obtenido es compatible con cierto valor poblacional.

Como ejemplo clásico de esa situación consideremos que se mide la tensión arterial diastólica (TAD) en una muestra de 36 diabéticos de 30 años y obtenemos *Media* = 90 mm de Hg y error estándar = 4. Sabemos que en esa región la TAD media de la población de sanos de 30 años es 74, y nos preguntamos si en los diabéticos la media poblacional estará aumentada, es decir, si será mayor que en la población de sanos. Los colectivos y medias que entran en juego son:

Población de sanos: TAD media poblacional = 74

Población de diabéticos: TAD media poblacional = ? (desconocida)

Muestra de diabéticos: TAD media muestral = 90

TS → *Hipótesis nula*: la TAD media poblacional en diabéticos es 74.

Valor P del test: $P = 0,00003$, o 3 por cien mil, es decir, si en los diabéticos la media poblacional fuera 74, solo en 3 de cada cien mil muestras de ese tamaño encontraríamos, por simple variabilidad en el muestreo, media muestral 90 o mayor. Es muy difícil que aparezca una muestra con media del orden de 90 si la media poblacional es 74. Que la muestra de

nuestro estudio tenga media 90 constituye una considerable evidencia a favor de que la media poblacional no es 74. Debe ser mayor.

A partir del dato muestral también se puede calcular el IC al 95% para la media en la población de diabéticos: $IC \rightarrow 82$ y 98 . Es decir, tenemos 95% de confianza en que la media poblacional de diabéticos sea un valor comprendido entre 82 y 98.

COMPARACIÓN DE DOS MUESTRAS: VALOR DE P GRANDE

Ahora veamos una nueva situación en la que la hipótesis nula no establece que la media poblacional de una variable es cierta cantidad, sino que dos medias poblacionales son iguales.

Obviamente la observación de las medias de muestras de esas dos poblaciones no nos permitirá, en ningún caso, conocer la diferencia exacta entre las medias poblacionales. El fundamento lógico es, como en el caso de una sola muestra y de la vida cotidiana, que cuanto más diferencia haya entre las dos medias muestrales más nos inclinaremos a pensar que hay diferencia entre las medias poblacionales, y por tanto, a rechazar la hipótesis nula que propone la igualdad de dichas medias poblacionales.

Supongamos que tenemos dos muestras de personas *alcohólicas*, 36 varones y 47 mujeres y a partir de las dos medias muestrales intentamos saber si las medias poblacionales son o no iguales en ambos sexos. Estos son los datos de TAD de las dos muestras:

	N	Media	Error estándar
Varones alcohólicos	36	90	4
Mujeres alcohólicas	47	83	3

Los colectivos y medias que entran en juego ahora son:

Población de mujeres alcohólicas: media poblacional = ?

Población de varones alcohólicos: media poblacional = ?

Muestra de mujeres alcohólicas: media muestral = 83

Muestra de varones alcohólicos: media muestral = 90

TS $\rightarrow$ *hipótesis nula: en alcohólicos la media de la TAD de la población de mujeres = media de TAD de la población de varones.*

En este caso la P del test es la probabilidad de que aparezcan, por simple variación del muestreo, unas medias muestrales tan distantes entre sí como las encontradas en nuestro estudio o aún más distantes, si las medias poblacionales son iguales.

Si, por ejemplo, hubiéramos encontrado una distancia entre las medias muestrales de 0,7, todos aceptaríamos que esa pequeña diferencia muestral es compatible con que no haya diferencia entre las medias poblacionales. Y, si por el contrario, aparece una distancia entre las medias muestrales de 40 unidades, habrá consenso en que esos datos constituyen fuerte evidencia contra la H_0 . Pero distancias menos extremas, no tan pequeñas ni tan grandes, no es claro si son o no compatibles con la H_0 y es entonces cuando el cálculo del valor P puede ayudarnos.

En nuestro ejemplo se encuentra una diferencia de $90 - 83 = 7$ mm de Hg a lo que corresponde un valor $P = 0,08$, es decir, si las medias poblacionales fueran iguales en ambos sexos, en 8 de cada cien estudios como este aparecería, por simple variaciones del muestreo, que la media muestral de varones aventaja a la de mujeres en 7 o más unidades. En otros 8 por cada cien estudios ocurriría lo inverso, es decir, que la media muestral de mujeres aventajaría a la de varones en 7 o más unidades, de modo que en 16 de cada 100 estudios se encontrará que las medias muestrales se diferencian entre sí (a favor de uno o de otro sexo) tanto como se han diferenciado en este estudio o más. Esto lo expresamos diciendo que es $P_{BILATERAL} = 0,16$. Es claro que esta diferencia entre medias muestrales es fácilmente compatible con la H_0 , y por tanto no constituye evidencia fuerte contra ella. Ese alto valor P no nos permite descartar que las medias *poblacionales* sean iguales en alcohólicos de ambos sexos. La diferencia observada entre las medias muestrales es compatible con que las medias poblacionales sean iguales (lo que no equivale a afirmar que son iguales).

Podemos calcular un IC dentro del cual es muy probable que esté el valor de esa diferencia entre las medias poblacionales. La media de varones menos la media de mujeres fue en estas muestras $90 - 83 = 7$ y el IC al 95% para la diferencia de las medias poblacionales es $7 \pm 10 = -3$ y 17 , es decir, tenemos confianza de 95% en que la diferencia de medias pobla-

cionales sea desde 3 a favor de las mujeres hasta 17 a favor de los varones. Ese intervalo incluye el cero, es decir, que ambas medias poblacionales pueden ser iguales, como ya habíamos concluido al hacer el TS.

COMPARACIÓN DE DOS MUESTRAS: VALOR DE P MUY PEQUEÑO

Supongamos ahora que se mide la TAD en otras dos muestras, mujeres y varones, afectos de *obesidad* y que los resultados son estos:

	N	Media	Error estándar
Varones obesos	36	103	4
Mujeres obesas	47	83	3

Los colectivos y medias que entran en juego ahora son:

Población de mujeres obesas: media poblacional = ?

Población de varones obesos: media poblacional = ?

Muestra de mujeres obesas: media muestral = 83

Muestra de varones obesos: media muestral = 103

TS → *Hipótesis nula: en obesos la TAD media poblacional en mujeres = TAD media poblacional en varones.*

Ahora tenemos una diferencia de TAD muestral de $103 - 83 = 20$ y el test nos da $P = 0,00003$, es decir, si las medias poblacionales fueran iguales en obesos de ambos sexos, solo en 3 de cada cien mil estudios como este aparecería, por simple variación del muestreo, que la media muestral de varones aventaja a la de mujeres en 20 o más unidades. En otras 3 por cada cien mil muestras ocurriría lo inverso, es decir, que la media de muestral de mujeres aventajaría a la de varones en 20 o más unidades, de modo que $P_{BILATERAL} = 0,00006$. Esta diferencia entre medias muestrales es difícilmente compatible con la H_0 y por tanto constituye evidencia fuerte contra ella. Ese pequeño valor P sugiere claramente que las medias poblacionales no son iguales en obesos de ambos sexos, sino que es mayor en varones.

COMPARACIÓN DE DOS MUESTRAS: VALOR DE P INTERMEDIOS

Supongamos ahora que se mide la TAD en otras dos muestras, mujeres y varones afectados de *hipertiroidismo* y estos son los resultados:

	N	Media	Error estándar
Varones hipertiroides	36	100	4
Mujeres hipertiroides	47	110	3

Ahora tenemos una diferencia muestral de $110 - 100 = 10$ unidades y al hacer el TS encontramos $P = 0,023$, es decir, si las medias poblacionales fueran iguales en ambos sexos, en 23 de cada mil estudios como este aparecería, por simples variaciones del muestreo, que la media muestral de mujeres aventaja a la de varones en 10 o más unidades. En otros 23 de esos mil estudios la media muestral de varones aventaja a la de mujeres en 10 o más unidades y la $P_{BILATERAL} = 0,046$. Esta diferencia entre medias muestrales no es fácilmente compatible con la H_0 , pero tampoco es extremadamente incompatible con ella, y por tanto no nos permite sacar una conclusión clara al respecto. Es posible que las medias poblacionales sean iguales en mujeres y varones con hipertiroidismo y nosotros hayamos encontrado una diferencia de medias muestrales de un orden de magnitud que aparece con poca frecuencia por las variaciones del muestreo. No podemos saber si las medias poblacionales son iguales y nos ha salido un tipo de muestra poco habitual o, por el contrario, lo que ocurre es que la media poblacional de mujeres es mayor que la de varones y ello produce esa diferencia en las muestras.

Recuerde que no se sale de ese estado de incertidumbre por decir que habiendo decidido considerar significativos los resultados con $P < 0,05$, este lo es, ni por decir que habiendo decidido considerar significativos los resultados con $P < 0,01$, este no lo es. Estas frases no añaden nada a la información de que disponemos y la realidad es que nuestros resultados no permiten tomar postura, pues son compatibles con que sea cierta la H_0 y también con que no lo sea. La comunidad científica no se inclinará a aceptar que hay diferencias entre las medias poblacionales hasta que no aparezcan más resultados que apunten en la misma dirección.

OTRAS HIPÓTESIS NULAS

Volvamos al ejemplo de los *hipertiroideos*:

	N	Media	Error estándar
Varones hipertiroideos	36	100	4
Mujeres hipertiroideas	47	110	3

En este ejemplo la evidencia contra la hipótesis nula que dice *que las medias poblacionales son iguales en varones y mujeres* es moderada. Podemos hacer tests para otras hipótesis «nulas», que propondrán que en las correspondientes *poblaciones* la diferencia entre mujeres y varones tendrá cierto valor.

- a) Por ejemplo, un endocrinólogo propone que en realidad la media poblacional en mujeres es **15** unidades mayor que la de varones. Ahora la hipótesis nula, H_0 , es: *media poblacional*_{mujeres} – *media poblacional*_{varones} = 15. Si eso fuera cierto y se hicieran muchos estudios como este, la cantidad «media muestral_{mujeres} – media muestral_{varones}» tomaría valores próximos a 15 en la mayoría de ellos. Queremos saber con cuánta frecuencia esa diferencia de medias muestrales es 10 (resultado obtenido en nuestras muestras) o menor. El test da valor $P = 0,16$. Si *media poblacional*_{mujeres} – *media poblacional*_{varones} = 15, en 16 de cada 100 estudios, aparecerá, por simple azar, que «media muestral_{mujeres} – media muestral_{varones}» es 10 o menor. En otros 16 estudios de cada 100, aparecerá, por simple azar, que «media muestral_{mujeres} – media muestral_{varones}» es 20¹ o mayor. Obtener una diferencia de medias muestrales del tipo de la encontrada es compatible con que la diferencia de medias poblacionales sea 15 y por ello esa hipótesis no se rechaza, sino que se acepta como posible.
- b) Otro endocrinólogo propone que en realidad la media poblacional en mujeres es **35** unidades mayor que la de varones. Ahora la hipótesis nula, H_0 , es: *media poblacional*_{mujeres} – *media poblacional*_{varones} = 35. Si eso fuera cierto y se hicieran muchos estudios como este, la cantidad «media muestral_{mujeres} – media muestral_{varones}»

¹ Es la situación «simétrica» a la obtenida en la muestra: 20 se diferencia de 15 lo mismo que 10, pero en el otro sentido.

tomaría valores próximos a 35 en la mayoría de ellos. Queremos saber con cuanta frecuencia esa diferencia de medias muestrales es 10 o menor. El test da $P = 0,0000003$. Si la *media poblacional mujeres* – *media poblacional varones* = 35, solo en 3 de cada 10 millones de estudios aparecerá, por simple azar, que «media muestral mujeres – media muestral varones» vale 10 o menos. En otros 3 estudios de cada 10 millones, aparecerá, por simple azar, que la «media muestral mujeres – media muestral varones» vale 60 o más². Obtener una diferencia de medias muestrales del tipo de la encontrada es muy difícilmente compatible con que la diferencia de medias poblacionales sea 35, y por ello esa hipótesis se rechaza. Creemos que la diferencia de medias poblacionales es menor de 35.

TEST DE SIGNIFICACIÓN E INTERVALO DE CONFIANZA PARA COMPARAR DOS PROPORCIONES

Consideremos el caso en el que decimos que cierto tipo de deporte (D) disminuye el riesgo de padecer infarto de miocardio (IM). Ahora la *hipótesis de trabajo* dice que la proporción de IM es menor en la población de deportistas que en la de sedentarios, si ambas son equiparables en las demás características.

Esa afirmación será válida y tendrá interés médico tanto si la diferencia es, por ejemplo, de 5 puntos (efecto moderado) como si es de 30 puntos (efecto importante). En ambos casos sería cierta la hipótesis de trabajo: «El ejercicio disminuye el riesgo de IM».

Para intentar llegar a conclusiones razonables al respecto mediremos la proporción de IM en dos muestras, una de deportistas y otra de sedentarios. Pero a partir de las proporciones *muestrales* no podemos saber cuánto menor es exactamente la proporción de IM en la *población* de deportistas que en la de sedentarios. La diferencia de las dos proporciones muestrales no nos permite saber cuánto es exactamente la diferencia de las proporciones poblacionales.

Veremos si las proporciones muestrales sugieren que las proporciones poblacionales *no son iguales*, lo que indicaría que el ejercicio hace

² Es la situación «simétrica» a la obtenida en la muestra: 60 se diferencia de 35 lo mismo que 10, pero en el otro sentido.

efecto, aunque no se especifique cuánto. Posteriormente, para hacer esa especificación usaremos el intervalo de confianza.

Téngase en cuenta que si la hipótesis de trabajo planteara, por ejemplo, que con ese deporte el porcentaje (%) de IM baja **20 puntos** respecto al % en sedentarios, sería imposible probar su veracidad exacta a través de los % encontrados en las dos muestras estudiadas. La diferencia de % entre las dos muestras no nos permite conocer con exactitud la diferencia de los % en las correspondientes poblaciones. Pero sí puede permitirnos, en algunos casos, saber que esos dos % poblacionales *no son iguales*.

El TS consiste en plantear la hipótesis nula que dice *que la diferencia en los % de IM entre la población de deportistas y de sedentarios es nula*, es decir, que hay el mismo % de IM en ambas poblaciones. Entonces se calcula la probabilidad de encontrar muestras con diferencias de % de IM como las encontradas o aún mayores, si en las poblaciones no hubiera diferencia. Este es el valor P del test.

Para afianzar mejor estas ideas sobre ejemplos concretos, supongamos que en una muestra de $N = 200$ deportistas de 70 años se encuentra que 6 hacen IM en el plazo de un año, mientras que en una muestra de 100 sedentarios lo hacen 7.

- Resumamos los datos en esta tabla:

	N	IM	% IM
Deportistas	200	6	3%
Sedentarios	100	7	7%

Ahora los colectivos y proporciones que entran en juego ahora son:

Población de deportistas: proporción poblacional de IM = ?

Población de sedentarios: proporción poblacional de IM = ?

Muestra de deportistas: proporción muestral = 0,03

Muestra de sedentarios: proporción muestral = 0,07

TS → Hipótesis nula: *proporción poblacional en deportistas = proporción poblacional en sedentarios*.

Al hacer el test se encuentra $P_{BILATERAL} = 0,19$, es decir, si realmente no hay diferencias entre ambos grupos (si el % de IM es el mismo en la

población de sedentarios que en la de deportistas) en 19 de cada 100 estudios como este se obtendrán muestras con diferencia de incidencia de IM como la encontrada en este estudio ($7\% - 3\% = 4\%$) o aún mayores. Por tanto, una diferencia de este tipo es compatible con que no haya diferencias poblacionales y no constituye una evidencia fuerte a favor de que la incidencia de IM sea menor en la población de deportistas, puesto que podrían ser iguales y encontrarse fácilmente una diferencia muestral de este tipo.

- Pero si en la muestra de 100 sedentarios se hubieran encontrado 23 IM, es decir, con estos datos:

	N	IM	% IM
Deportistas	200	6	3%
Sedentarios	100	23	23%

Al hacer el test se encuentra $P_{BILATERAL} = 0,000002$, es decir, si realmente no hay diferencias entre ambos grupos, en 2 de cada millón de estudios como este se obtendrán muestras con diferencia de incidencia de IM tan grande como la encontrada en este estudio, $23\% - 3\% = 20\%$, o aún mayor. Por tanto, una diferencia de este tipo es muy difícil que aparezca si no hay diferencias poblacionales, y el hecho de que haya aparecido en las muestras constituye evidencia muy fuerte a favor de que la incidencia de IM no es igual en ambas poblaciones. Por lo que concluiríamos que es menor en la población de deportistas.

- Y si en la muestra de sedentarios se hubieran encontrado 9 IM, es decir, con estos datos:

	N	IM	% IM
Deportistas	200	6	3%
Sedentarios	100	9	9%

Al hacer el test se encuentra $P_{BILATERAL} = 0,04$, es decir, si realmente no hay diferencias entre ambos grupos, en 4 de cada 100 estudios como este se obtendrán muestras con diferencia de incidencia de IM tan grande

como la encontrada en este estudio, $9\% - 3\% = 6\%$, o aún mayor. Una diferencia de este tipo no es muy difícil, pero tampoco es muy fácil, que aparezca si no hay diferencias poblacionales. Estos resultados constituyen una moderada evidencia en contra de la H_0 , y a favor de que la incidencia de IM es menor en la población de deportistas, pero no podemos descartar la posibilidad de que sea cierta dicha hipótesis y hayamos encontrado un tipo de muestra que es poco frecuente obtener a partir de poblaciones en las que no hay diferencias de incidencia de IM.

COMPRUEBE SU NIVEL DE CONOCIMIENTOS: ENCUESTA DE AUTOEVALUACIÓN

En el Apéndice 2 encontrará una encuesta de autoevaluación para este capítulo, que le ayudará a evaluar en qué medida tiene claras sus ideas en este tema.

No afirmar la hipótesis nula

Uno de los errores más frecuentes en el análisis estadístico de datos es confundir el «no rechazar» una hipótesis con «afirmar que es cierta».

Ya sabemos que si el valor P del test es grande aceptamos que la H_0 , que dice que no hay efecto en la población, *puede ser* cierta porque los datos muestrales son claramente compatibles con ella, pero *no afirmamos* que sea cierta, porque los datos muestrales también son compatibles con otras hipótesis. Decimos que «el resultado es no significativo», indicando que el efecto encontrado en la muestra no supone fuerte evidencia a favor de que exista ese tipo de efecto en la población.

El error está en creer que «no significativo» indica que no existe en la población general ese tipo de efecto.

La experiencia muestra que es difícil para muchos investigadores entender que cuando los datos observados son compatibles con la hipótesis el «no rechazar» esta, no implica «afirmarla», sino considerar que puede ser cierta.

Son dos conceptos muy diferentes que todas las personas, con y sin formación académica, distinguen perfectamente en la vida diaria. Cada día usamos espontánea y correctamente este razonamiento cientos de veces. Para no errar en este aspecto de la Inferencia Estadística solo tenemos que aplicar la misma lógica que en la vida común. A continuación vemos varios ejemplos de los que ocurren cada día y de los específicos de la investigación.

EJEMPLOS DE LA VIDA DIARIA

a) Los ministros no suelen viajar en el metro

Dos amigos van en el metro y uno de ellos comenta: «Aquel señor en el fondo del vagón parece el ministro de Comercio». El otro responde: «Desde aquí casi no puedo verlo, pero es prácticamente seguro que no es el ministro, porque si lo fuera no estaría viajando en metro». Es decir, el dato observado (aquel señor está viajando en el metro) no es compatible con la hipótesis (aquel señor es el ministro).

Al salir a la superficie ven pasar un coche oficial con escolta y uno de los amigos comenta: «El señor que va en ese coche se parece al ministro». El otro responde: «Desde aquí casi no puedo verlo, pero *es posible* que lo sea, puesto que si lo fuera es normal que viaje en coche oficial». Es decir, el dato observado (aquel señor está viajando en coche oficial) es compatible con la hipótesis (aquel señor es el ministro).

Observe que en ninguno de los dos casos se afirmó que se tratara del ministro. Lo único que se dijo es que el viajar en metro es un dato en contra de que sea el ministro, mientras que el viajar en coche oficial no es un dato en contra de que sea el ministro. Y debe estar claro que el viajar en coche oficial no es un dato a favor de que sea el ministro, hay otros muchos personajes que viajan en coche oficial. Lo que ocurre es que no es un dato en contra.

Debe tener muy clara esta situación porque al elaborar las conclusiones de una investigación se usa exactamente el mismo proceso lógico.

b) El asesino no puede estar muy lejos

Para insistir en que aceptar que una hipótesis puede ser cierta no es afirmar que lo sea retomemos otro ejemplo ya conocido. En una señorial mansión se comete un crimen a las 12:00 h. Naturalmente, Bautista, el mayordomo, es uno de los sospechosos. Consideramos la hipótesis, H_0 : *Bautista es el asesino*.

Ocurre que nadie ha visto cometerse el asesinato pero algunos testigos muy fiables dan información sobre la ubicación de Bautista *a las 12:15 h.*

1. Si Bautista fue visto a las 12:15 h en una ciudad a 900 km de la casa, usted concluye que Bautista **no es** el asesino, es decir, rechaza H_0 .
2. Pero si Bautista fue visto a las 12:15 h en el portal de la mansión: usted *no* concluye que Bautista es el asesino, sino que **puede serlo**. Usted acepta la H_0 como posible, pero no afirma que sea cierta.

En general, valores grandes de P (equivalentes a haber visto al sospechoso en el lugar y a la hora del crimen) indican que los datos no ofrecen evidencia contra la hipótesis nula, pero no permiten afirmar que sea cierta, solo nos dicen que el efecto encontrado en la muestra es compatible con la hipótesis nula, además de serlo con otras hipótesis.

c) La lógica del diagnóstico médico

Rara vez los médicos emiten un diagnóstico en términos de certeza. Por el contrario, lo habitual es que utilicen expresiones como: «Los hechos observados en el paciente son compatibles con tal enfermedad y permiten descartar tal otra» y en ese acto están usando el mismo proceso lógico que en los tests de significación. Cuando el médico dice «Los datos del paciente son compatibles con la enfermedad 'A', no está asegurando que tenga esa enfermedad, solo dice que puede que la tenga. Del mismo modo, un valor P grande nos dice que el resultado muestral es compatible con que sea cierta la hipótesis nula, no que lo sea necesariamente.

Por otra parte, cuando el médico dice «Los datos del paciente permiten descartar la enfermedad 'B' », es que tiene mucha seguridad en ello, porque encuentra aspectos que son claramente incompatibles con esa enfermedad. Del mismo modo, un valor de P muy pequeño nos dice que el resultado muestral es incompatible con la hipótesis nula planteada en un test de significación.

EJEMPLOS DE INFERENCIA ESTADÍSTICA

En la farmacia de un hospital se compra a la mitad de su precio habitual un gran envase con 100.000 pastillas, porque el vendedor nos asegura que están deterioradas el 20% de ellas (el defecto es claramente visible y no hay riesgo de confundirlas con las correctas). Para ver si realmente

están deterioradas el 20% tomamos una muestra de 50 pastillas. Si lo que nos dijo el vendedor es cierto, esperamos encontrar unas 10 deterioradas (10 es el 20% de 50). Si encontramos un número ligeramente superior a 10 no nos parecerá especialmente sospechoso. Y si el número de deterioradas es mucho mayor de 10, sospecharemos que el porcentaje de ellas en el envase es superior a 20%. Cuanto más pastillas deterioradas aparezcan en la muestra, más sospechamos que en el envase hay más de 20% deterioradas.

Pero si encontramos un número de deterioradas próximo a 10, no aseguramos que en el envase haya un 20%, solo decimos que el dato no constituye evidencia contra la hipótesis que dice que son un 20%. Por ejemplo, si en la muestra aparecen 12 deterioradas (24%) decimos que en el envase *pueden ser* deterioradas el 20%, pues ese dato no es argumento fuerte contra esa hipótesis. Al calcular el IC (al 99% de confianza) para el % del envase se encuentra: 11% y 43%, que incluye el 20%.

Incluso, si en la muestra aparecen 10 deterioradas (justamente el 20%), decimos que en el envase *pueden ser* deterioradas el 20%, es decir, este dato no es argumento contra esa hipótesis. Al calcular el IC (confianza 99%) encontramos que muy probablemente el % de deterioradas en el envase está entre 8% y 38%. Repitamos una vez más que un % muestral próximo al 20%, o incluso igual al 20%, no es evidencia a favor de que el % del envase sea exactamente el 20%, solo es falta de evidencia a favor de que el % en el envase es distinto del 20%.

EJEMPLOS DE LA INVESTIGACIÓN

Haremos énfasis en esta idea utilizando de nuevo el ejemplo del capítulo 3.

Para estudiar el posible efecto anticancerígeno (AC) de 2 productos, «A», y «D», trabajaremos con ratas de una cepa en la que el 60% de ellas desarrollan cáncer de cérvix espontáneamente.

Probaremos cada fármaco en 40 ratas. Si no es AC esperamos que unas 24 hagan cáncer (24 es el 60% de 40). Cuanto menor sea el número de ratas que desarrollan cáncer más nos inclinaremos a pensar que hay efecto AC. He aquí los resultados, el valor P del test y los intervalos de confianza.

Fármaco	Núm. de ratas con cáncer	% de ratas con cáncer	Valor P IC al 99%
A	8	20%	0,000003 7%-41%
D	23	57,5%	0,436 36%-77%

Conclusiones para «A»

Es prácticamente seguro que «A» es AC, pues es muy difícil que aparezca ese tipo de muestra con 8 cánceres, si «A» no fuera AC. (Si A no fuera AC, solo en 3 muestras cada millón aparecerían 8 o menos cánceres). Además, calculado el intervalo de confianza al 99% encontramos que si diéramos «A» a toda la población, la proporción de cánceres estará entre 7% y 41%, claramente por debajo del 60%.

Conclusiones para «D»

Supongamos que el resultado para este producto era esperado con especial expectación porque la gran mayoría de los profesionales insistían en que «D» era totalmente inútil y no tenía sentido hacer este experimento. Cuando finalmente llegan los resultados (57,5% de cánceres en la muestra tratada), la mayoría de los observadores dicen:

«Estos datos prueban que «D» no es anticancerígeno. En la muestra tratada con «D» se obtuvo un número de cánceres muy próximo al esperado cuando no se da producto alguno. Es imposible encontrar un resultado más desalentador. El fracaso es total».

Pero esa conclusión es arbitraria, los datos no la avalan. Pues aunque el resultado es compatible con que «D» no sea AC, también lo es con que «D» sea un buen AC. En el intervalo de confianza al 99%, vemos que dando «D» la proporción real de cánceres estará entre 36% y 77%. Por tanto «D» puede que disminuya el % de cánceres hasta en 24 puntos ($60 - 36 = 24$), lo cual representaría un efecto muy notable. Pero también puede que aumente el % de cánceres hasta en 17 puntos ($77 - 60$).

= 17)¹. Y entre esas dos posibilidades está la de que no modifique en ningún sentido ese %.

NO HAY SIMETRÍA

Cuando el resultado es difícilmente compatible con la hipótesis, esta se rechazará con gran seguridad. Pero no hay «simetría» en este tipo de razonamiento. Si los datos son compatibles con la hipótesis, no podemos afirmar que sea cierta, solo diremos que *puede* serlo.

La experiencia muestra que es difícil para muchos biólogos entender que cuando los datos observados son compatibles con la hipótesis, el «no rechazar» esta no implica «afirmarla».

Incluso algunos estadísticos tienden a confundir estas dos posturas lógicas, porque habiendo aprendido el mecanismo de los tests estadísticos enfocados, como Neymann y Pearson lo hicieron, a la *Toma de Decisiones*, creen que en Biología se investiga para tomar decisiones fácticas.

En general, es un craso error confundir la finalidad de la investigación científica, destinada a valorar la evidencia que se va acumulando a favor de cierta hipótesis, con la finalidad de la *Toma de Decisiones*, en la que a partir de lo observado en la muestra se decide una u otra acción.

Recuerde que «El propósito central de un experimento no es precipitar la toma de decisiones, sino propiciar un reajuste en el grado de confianza que uno tiene en la veracidad de cierta hipótesis... la tarea del científico no es prescribir acciones sino establecer convicciones razonables.» L. C. Silva.

Por eso, frases como «decidimos que la hipótesis nula es cierta, ya que hemos encontrado $P > 0,05$ » no tienen sentido en la investigación científica, en la que no se trata de decidir, sino de valorar en qué medida los resultados obtenidos constituyen evidencia a favor de una hipótesis.

¹ Por supuesto, hay una probabilidad de 0,005 de que el porcentaje poblacional con «D» sea inferior a 36%, y también una probabilidad de 0,005 de que sea superior a 77%.

COMPRUEBE SU NIVEL DE CONOCIMIENTOS: ENCUESTA DE AUTOEVALUACIÓN

En el Apéndice 2 encontrará una encuesta de autoevaluación para este capítulo, que le ayudará a evaluar en qué medida tiene claras sus ideas en este tema.

Capítulo 11

La falsa frontera del 5%

En este capítulo se explica que la frontera del 5% es arbitraria y en la mayoría de los casos no procede referirse a ella.

Lamentablemente, la mayoría de los profesionales de la investigación están expuestos a cometer errores serios al elaborar las conclusiones de sus trabajos experimentales por creer que el valor $P = 0,05$ constituye una frontera decisiva con propiedades especiales.

Algunos editores de revistas científicas y muchos *referees* dan esta norma: «Los resultados se declararán *significativos* si es $P < 0,05$ y “*no significativos*” si es $P > 0,05$ ». Este convenio puede ser un resumen útil en algunas situaciones, pero frecuentemente es mal interpretado y lleva a muchos investigadores a cometer fallos notables.

El error está en creer que «significativo» (porque es $P < 0,05$) indica que tenemos la seguridad de que el tipo de efecto encontrado en la muestra existe realmente en la población. Y que «no significativo» (porque es $P > 0,05$) implica que ese tipo de efecto no existe en la población general y se ha producido en la muestra por azar.

Estos errores se evitan entendiendo lo que el valor P indica, lo cual ya sabemos que está al alcance de todo investigador, independientemente de sus conocimientos matemáticos. Repitamos una vez más que calcular el valor P en cada test es una cuestión matemática, pero interpretarlo correctamente es una cuestión de lógica común que todos podemos entender.

NO HAY UN VALOR DE P QUE SEA FRONTERA

Recordemos que valores de P muy pequeños nos llevan a pensar que el tipo de efecto encontrado en la muestra también se da en la población, es decir, a rechazar la H_0 que dice que no hay tal tipo de efecto en la población, porque los datos experimentales son difícilmente compatibles con ella. Decimos que el resultado es «significativo». Si P es grande aceptamos que la H_0 puede ser cierta porque los datos muestrales son claramente compatibles con ella, pero no afirmamos que sea cierta, es decir, pensamos que el efecto de la muestra pudo haber aparecido por azar, aunque también es posible que haya ese tipo de efecto en la población. Decimos que el resultado es «no significativo», indicando que no constituye fuerte evidencia a favor de que exista ese tipo de efecto en la población.

¿Y qué valores de P debemos considerar «pequeños» y cuáles «grandes»? ¿Qué valor de P marca la frontera en ese sentido?

El investigador debe saber que *no* hay un valor P que marque una frontera natural entre estas dos posturas, sino que es una escala continua. Cuanto menor sea el valor P más evidencia aporta contra la H_0 y a favor de que ese tipo de efecto ocurre en la población general, pero no hay un punto de separación. Entender esto no debería ser problema para ningún lector, porque en su *vida normal* se encuentra a diario con este tipo de situaciones y en todas ellas entiende perfectamente que ciertas magnitudes *varían de modo continuo sin que haya un punto de separación* que delimite dos zonas. En el siguiente apartado vemos algunos ejemplos muy obvios.

EN LAS VARIABLES CONTINUAS NO HAY UN PUNTO DE CORTE

En la vida común manejamos constantemente variables continuas sin intentar reducirlas a un «sí» o «no», que resultaría totalmente artificial.

a) El lactante y su abuela tienen edad «equivalente»

Por ejemplo, nos referimos a la edad de las personas dando sus años, no diciendo si su edad está a uno u otro lado de cierta cantidad. Imagine un hospital donde para cada paciente no se conociera su edad, tan solo si es o no mayor de 40 años. Ello implicaría situar en el mismo grupo de

edad a una persona de 41 años y otra de 93 y en diferente grupo a una que ha cumplido 41 ayer de otra que los cumplirá el próximo mes. Serían semejantes en edad el lactante y su abuela de 40 años. Y si para evitar que el lactante y su abuela estén en el mismo grupo ponemos la edad de corte en otro valor, aparecen otras distorsiones tan inapropiadas como las primeras. Por ejemplo, si ponemos la frontera en los 30 años, tendría edad equivalente el campeón olímpico de 31 y su abuelo de 92. El problema no está en el valor elegido como frontera, sino en intentar establecer un valor frontera que reduzca a dos categorías una variable que toma muchos valores distintos. Tal artificio es inaceptable, tanto en el contexto hospitalario mencionado como en la vida ordinaria.

Igualmente impropio es considerar dos resultados experimentales muy distintos porque uno da valor P del test menor de 0,05 y el otro supera esa cantidad. O porque están a distinto lado de $P = 0,01$ o de cualquier otro valor convenido.

b) El anoréxico de 45 kg y su saludable vecino de 79 kg tienen peso «equivalente»

Entre los miles de ejemplos que en este mismo sentido pueden tomarse de la vida cotidiana consideremos ahora el peso de las personas. Supongamos un varón de 168 cm de estatura. Imagine que en su historia clínica no se reflejara su peso, tan solo si pesa o no más de 80 kg, por ejemplo. Ello implicaría considerar equivalentes en peso a uno de 81 kg y otro de 120 kg y diferentes a uno que pesa 79,9 de otro que pesa 80,1. Y serían semejantes el anoréxico con 45 kg y el saludable de 79,9 kg. También esta dicotomización parece inaceptable tanto en el hospital como en la vida ordinaria.

c) ¿Qué riesgo de muerte aceptaría el paciente?

Supongamos que para los pacientes con cierta dolencia crónica invalidante se desarrolla una nueva técnica quirúrgica que muy probablemente elimine sus problemas y les devuelva a la normalidad, pero hay cierta probabilidad —diferente para cada paciente— de morir en el acto quirúrgico.

Si para un paciente el riesgo de muerte en la operación es, por ejemplo, $R_M = 0,90$, probablemente optará por no operarse. Otro paciente con $R_M = 0,0000001$ (se muere 1 cada diez millones operados) optará por operarse.

En general, con R_M grandes se elige no operación y con R_M muy pequeño se elige operación. ¿Pero dónde está la separación entre estas dos opciones? ¿Qué R_M separa los valores que llevan a elegir operación de los que llevan a no hacerlo? Es obvio que no hay una cantidad que marque el límite.

Imaginemos un paciente que está en duda porque su R_M es 0,052 (se mueren 5,2 % de los operados) y tiempo después se le dice que la técnica ha mejorado, de modo que la probabilidad de muerte ya no es 0,052 sino 0,049. ¿Cree usted que las dudas del paciente y su eventual decisión cambiarán mucho por haber pasado el R_M de 5,2% a 4,9%?

Por la misma razón al hacer inferencia en investigación, un valor $P = 0,052$ no puede llevarnos a distinta conclusión lógica que un valor $P = 0,049$.

INTENTO DE APLICAR LOS MODOS DE LA TOMA DE DECISIONES AL PROCESO DE ADQUISICIÓN DE CONOCIMIENTOS

Para los investigadores acostumbrados durante muchos años a usar la regla del 5% es difícil asumir que en el proceso de formarse opinión nuestra mente no usa puntos de corte. Su argumentación es:

«Decir que la H_0 se rechaza cuando el valor P es “muy pequeño” es muy ambiguo y cabe exigir más precisión en ese aspecto. ¿Qué valores de P son “muy pequeños” y cuáles no lo son? Si, por ejemplo, con $P = 0,000001$ rechazamos la H_0 , y con $P = 0,30$ no la rechazamos, debe haber algún punto en el que comienza la “región de rechazo”. Lo adecuado es convenir un valor para ese punto de comienzo, de modo que se rechaza H_0 cuando P no alcanza ese valor y se acepta cuando P lo supera».

Veamos detalladamente ese argumento.

En efecto, como norma a seguir para optar entre dos acciones, la expresión «valores de P muy pequeños» sería inaceptablemente ambigua. Pero al decir que «la H_0 se rechaza cuando el valor P es muy pequeño» no estamos proponiendo una norma de acción, sino relatando cómo funciona nuestra mente. Ante un resultado muy difícilmente compatible con la H_0 pensamos que no es cierta y eso es una obligación lógica a la que no podemos sustraernos. Y también forma parte de la naturaleza humana el que nuestra creencia en la falsedad de la H_0 *crezca progresiva-*

mente al disminuir el valor P, sin que haya una cantidad que separe los valores de rechazo de los valores de aceptación.

Esto no es algo específico de la Inferencia Estadística, sino común a todos los órdenes de la vida. Hay muchas magnitudes con regiones de valores que llevan a una opinión bien definida, sin que esas regiones empiecen en una cifra determinada, sino que están separadas de otras regiones por zonas de *transición gradual* en las que el observador no puede formarse opinión firme, pero su postura se hace más sólida cuanto más extremo se hace el valor.

Hay cientos, miles, de ejemplos en este sentido y es pertinente insistir en ello para que quede claro que el mecanismo mental de los TS es propio de los humanos en todas sus facetas, y no implica un «fallo» o carencia, sino que es el más acorde a la realidad. Considere la H_0 que dice que al buzo que se sumergió con aire para dos horas no le ha ocurrido ningún percance. Veamos nuestra opinión según el tiempo que llevamos esperando a que emerja. Si el tiempo es de 1 hora y 40' o 1 h y 50' aceptamos la H_0 , si el tiempo es 4 horas, la rechazamos categóricamente, y si el tiempo es 2 h y 5' empezamos a dudar de la hipótesis y a medida que el tiempo de espera va creciendo nuestra duda se acrecienta *progresivamente* hasta llegar al convencimiento de que ha sufrido un accidente (rechazo de la hipótesis inicial).

Decimos que esta actitud es la más acorde con la naturaleza porque refleja el hecho empírico de que la mayoría de los retrasos de 5' acaban en que no hubo accidente previo, y el % de casos con accidente es mayor en los retrasos de 15' y mayor en los de 20'.... y es prácticamente el 100% en los que duraron 4 horas. Nuestra creciente tendencia a rechazar la H_0 a medida que crece el tiempo de espera refleja el creciente % de casos en que hubo accidente cuanto mayor fue esa espera. Y es obvio que no hay un tiempo tal que toda demora por encima de él se asocie con accidente y toda demora por abajo de él se asocie con no accidente. No hay un tiempo frontera¹.

Cada día, en la vida cotidiana, usamos implícitamente muchas veces el mecanismo lógico de los TS. Nuestro convencimiento de que una hipótesis es falsa, aumenta gradualmente a medida que lo observado se aleja más y más de lo esperado bajo la hipótesis, sin que haya un punto de corte.

¹ Y, una vez más, si hay que decidir entre ejecutar una acción u otra, es pertinente poner un punto de corte. El club submarinista podría establecer esta norma: cuando la demora llega a 2 h y 15' se inicia la inmersión de rescate. Ello no implica que ese tiempo marque una frontera conceptual. Solo establece un criterio de acción necesario en la práctica.

La hipótesis que dice que un paciente no padece una infección se rechaza tajantemente si su temperatura es 40° , no se rechaza si es $36,5^{\circ}$ y la sospecha de que hay infección va subiendo a medida que la temperatura va tomando valores de 37° , $37,2^{\circ}$, $37,5^{\circ}$...

La hipótesis que dice que una moneda es equilibrada (igual proporción de caras que de cruces) se rechaza tajantemente si al lanzarla 80 veces salen 77 caras, no se rechaza si salen 41 y la sospecha de que está sesgada es tanto mayor cuanto más caras hayan aparecido. Pero no hay un número tal que si la cantidad de caras lo supera nos hace estar seguros de que la moneda está sesgada y si no lo alcanza nos hace estar seguros de que es equilibrada.

En los tests estadísticos nuestro intelecto no llega a distinta opinión porque el valor P esté a uno u otro lado de cierta cantidad.

EJEMPLOS NUMÉRICOS

Haremos énfasis en esta idea usando de nuevo el ejemplo del Capítulo 3. Para estudiar el posible efecto anticancerígeno (AC) de 4 productos, «A», «B», «C» y «D», trabajaremos con ratas de una cepa en la que el 60% de ellas desarrollan cáncer espontáneamente.

Probaremos cada fármaco en **40** ratas. Si no es AC esperamos que unas **24** hagan cáncer (24 es el 60% de 40). Cuanto menor sea el número de ratas que desarrollan cáncer más nos inclinaremos a pensar que hay efecto AC. Cuando el resultado no es concluyente se calcula el «valor P del test». He aquí los resultados, el valor P del test y los intervalos de confianza.

Fármaco	Núm. de ratas con cáncer	% de ratas con cáncer	Valor P IC al 99%
A	8	20%	0,000003 7%-41%
B	18	45%	0,039 25%-66%
C	19	47,5%	0,074 27%-68%
D	23	57,5%	0,436 36%-77%

Veamos las conclusiones obtenidas creyendo que el valor $P = 0,05$ marca una diferencia definitiva:

«“A” y “B” son anticancerígenos ($P < 0,05$) por el contrario “C” y “D” no lo son ($P > 0,05$)».

Las conclusiones razonables son:

«El fármaco “A” parece ser un potente AC, mientras que “B, C, D” pueden serlo o no serlo».

Veamos más detenidamente las conclusiones para cada fármaco.

Conclusiones para «A»

Es prácticamente seguro que «A» es AC, pues es muy difícil que aparezcan 8 cánceres o menos si «A» no fuera AC.

Además, calculado el intervalo de confianza al 99% encontramos que con A la verdadera proporción de cánceres estará entre 7% y 41%, claramente por debajo del 60%.

Conclusiones para «B»

«B» puede que sea AC, pero también puede que no lo sea. Con este resultado es imposible pronunciarse.

No es muy fácil, pero tampoco muy difícil, encontrar un resultado muestral de ese orden si realmente «B» no fuera AC (3,9% es la probabilidad de encontrar solo 45% o menos cánceres por casualidad). Podría ser que la disminución de cánceres encontrados en la muestra ($60 - 45 = 15$ puntos) haya sido puro azar del muestreo.

Además, el intervalo de confianza al 99% nos dice que muy probablemente la proporción real de cánceres con «B» estará entre 25% y 66%. Por tanto puede que «B» disminuya el % de cánceres en 35 puntos ($60 - 25 = 35$) o algo más, puede que no lo modifique (60%) y también puede que lo aumente en 6 puntos ($66 - 60 = 6$) o algo más.

El intervalo de confianza se podría haber calculado, por supuesto, a cualquier otro nivel, en cuyo caso, los límites serían otros. Esto no impli-

ca contradicción alguna con los correspondientes al 99%. Ninguno de estos límites tiene carácter de frontera determinante, sino que orientan del tipo de valores que es más probable tenga el parámetro poblacional. Esta orientación es en términos de probabilidad y, en ningún caso, permite excluir definitivamente valores que estén fuera de los límites.

Conclusiones para «C»

Puede que «C» sea AC, pero también puede que no lo sea. Con este resultado es imposible pronunciarse.

No es muy fácil, pero tampoco muy difícil encontrar ese resultado muestral si realmente «C» no fuera AC. (7,4% es la probabilidad de encontrar solo 47,5% o menos casos de cánceres por casualidad). La disminución de cánceres encontrados en la muestra ($60 - 47,5 = 12,5$) podría ser puro azar del muestreo.

Además, calculado el intervalo de confianza al 99% encontramos que con «C» la proporción real de cánceres estará entre 27% y 68%. Por tanto, puede que «C» disminuya el % de cánceres en 33 puntos ($60 - 27 = 33$) o algo más, puede que no lo modifique y también puede que lo aumente en 8 puntos ($68 - 60 = 8$) o algo más.

Las conclusiones para «D» las vimos más detalladamente en el capítulo anterior.

Vemos que solo respecto a «A» podemos hacer una afirmación bastante contundente. Respecto a los otros tres fármacos no podemos tomar postura. Los resultados son compatibles con que sean y con que no sean AC.

Esta incertidumbre es inherente a estos resultados y el investigador debe asumirlo así. El único modo de salir de esa incertidumbre es ampliar la investigación a más individuos y mientras eso no sea posible, no hay más opción que aceptarla. Es inútil intentar disfrazar la incertidumbre con afirmaciones de apariencia contundente pero carentes de contenido, tales como «el resultado es estadísticamente significativo» o «el resultado no es estadísticamente significativo».

Reconociendo que no se puede tomar postura cuando no se puede, se evitan afirmaciones no justificadas y se sustituyen por información veraz que el lector maduro sabrá valorar adecuadamente, no sintiéndose el autor del trabajo ni sus lectores obligados a decantarse a favor o en contra de una hipótesis cuando los resultados no lo permiten.

COMPRUEBE SU NIVEL DE CONOCIMIENTOS: ENCUESTA DE AUTOEVALUACIÓN

En el Apéndice 2 encontrará una encuesta de autoevaluación para este capítulo, que le ayudará a evaluar en qué medida tiene claras sus ideas en este tema.

El origen del malentendido: pensar *versus* decidir

Pretender basar las conclusiones de una investigación en que el valor P esté a uno u otro lado de cierta presunta barrera decidida por el investigador (5% o 1% o cualquier otra cantidad) lleva a situaciones absurdas dignas del más puro surrealismo. Considere el caso de un estudio en el que se indaga el posible efecto anticancerígeno (AC) del producto «B»: en la muestra estudiada aparece un moderado efecto AC y el test estadístico da $P = 0,04$.

El Doctor Vargas dice:

«Había decidido considerar el resultado ‘estadísticamente significativo’ si salía $P < 0,05$, y por ello concluyo que ‘B’ es AC».

El Doctor. Llosa dice:

«Había decidido considerar el resultado ‘estadísticamente significativo’ si salía $P < 0,01$, y por ello concluyo que ‘B’ no es AC».

Es obvio que los dos no pueden estar en lo cierto. ¿Cuál de los dos se equivoca? En realidad ambos yerran al plantear la conclusión del trabajo en esos términos.

En primer lugar, que «B» sea o no sea AC no puede depender de la «decisión» de quien comenta el resultado. Depende de leyes químicas y biológicas que ignoramos en gran parte y sobre las que no tenemos ninguna capacidad de decisión. Al investigador no le compete «decidir» cómo es la naturaleza, sino «intentar saber» cómo es, «intentar llegar a la convicción» de que una hipótesis es cierta o es falsa.

En segundo lugar, con ese resultado no se puede saber si «B» es o no es AC, ya que el dato obtenido es compatible con ambas opciones. Ambos doctores deberían mencionar esta incertidumbre en sus conclusiones, en vez de hacer afirmaciones tan rotundas como gratuitas.

En resumen, en ningún caso está en la mano del investigador «decidir» cómo es la naturaleza y en este caso ni siquiera lo sabe.

Hagamos una reflexión cuidadosa sobre el término «decidir» que nos ayude a usarlo con más precisión.

«DECISIONES» IMPOSIBLES

En el lenguaje común usamos el verbo «decidir» con una laxitud que no crea confusión, pero al trasladar esa ambigüedad al lenguaje científico puede crear problemas. Veamos estas frases, habituales en la conversación cotidiana:

«... el cielo estaba parcialmente nublado y yo dudaba entre sacar o no el paraguas, pero finalmente *decidí* que luciría el sol y dejé el paraguas en casa».

«... no podíamos saber si aquel hombre que nos pedía alojamiento era de verdad un viajero extraviado o un ladrón malintencionado, pero finalmente *decidimos* que sería una persona honesta y le invitamos a pasar».

«... en el examen de geografía me preguntaron cuál era el río más largo del mundo. Yo dudaba entre el Nilo y el Amazonas, pero finalmente *decidí* que era el Nilo y así lo escribí».

A primera vista no hay nada confuso en estas frases y todos los lectores reciben con ellas la información que sus autores querían transmitirles. Pero contienen una ambigüedad que aunque en ellas no supone problema, al ser llevada a las conclusiones de los trabajos de investigación genera confusión importante. Escribamos de nuevo la parte que puede generar conflicto:

«... *decidí* que las nubes se irían y dejé el paraguas en casa».

«... *decidimos* que sería una persona honesta y le invitamos a pasar».

«... *decidí* que el río más largo del mundo era el Nilo y así lo escribí».

Podría parecer que los personajes de estas frases tiene poderes extraordinarios que les permiten *decidir* la climatología, las intenciones de un desconocido y la longitud de los ríos. Pero es obvio que ellos no

deciden el clima que va a hacer, solo deciden si sacan o no el paraguas. No deciden las intenciones de un visitante, solo deciden si le invitan a pasar o le cierran la puerta. No deciden la longitud de los grandes ríos, solo deciden la respuesta que escriben en el examen.

DIFERENCIA ENTRE «CONOCER» Y «ACTUAR»

Para aclarar el uso del término «decidir»¹ tenemos que subrayar la diferencia entre *conocer* y *actuar*, distinguiendo claramente estas dos situaciones:

- a) Formarse opinión sobre un tema → es una cuestión de «conocer».
- b) Ejecutar una acción → es una cuestión de «actuar».

Y tomando conciencia de que el término «decidir» tiene distinto significado en cada una de estos contextos. Veámoslo en ejemplos de la vida común.

Ejemplo: siguiendo las huellas del ladrón

El comisario Zaj quiere recuperar el millonario botín que los ladrones han escondido en una de dos grandes minas abandonadas, la del Norte y la del Sur, y solo tiene recursos espeleológicos para entrar en una de ellas. Inspeccionando cuidadosamente la entrada de ambas descubre signos de actividad reciente en la del Norte. En lenguaje coloquial es habitual describir el pensamiento y la acción de Zaj con este frase:

A la vista de esta información Zaj decide que el botín está en la Norte y decide entrar en ella.

Observe que el verbo «decidir» está propiamente usado en la segunda parte de la frase, pues el entrar en una u otra de las dos minas es algo que realmente decide Zaj. El «decide» de la primera parte de la frase debe ser aclarado, pues es obvio que la ubicación actual del botín la decidieron los ladrones en su momento (sin la participación de Zaj) y lo que el comisario hace ahora no es «decidir» que el botín esté en la mina Norte,

¹ Y ver lo inapropiado de frases como «Habíamos decidido considerar el resultado estadísticamente significativo si salía $P < 0,05$, y por ello concluimos que 'B' es AC».

sino *creer, pensar, tener el convencimiento de* que está allí, pues es la hipótesis más compatible con los datos observados.

En el lenguaje coloquial usamos el término «decidir» como sustituto de «estar convencido de», y ello no crea confusión porque se entiende claramente lo que se quiere decir. Pero esa misma laxitud puede dar lugar a confusiones notables en el lenguaje científico.

La frase adecuada para esa situación sería:

A la vista de esta información Zaj, está convencido de (o «piensa» o «cree») que el botín está en la Norte y decide entrar en ella.

El siguiente año se le presenta el mismo problema e inspeccionando la entrada de las dos minas descubre signos inequívocos de actividad reciente en ambas, de manera que esa información no le permite saber en qué mina está el botín. Entonces decide entrar en una de ellas, sabiendo que puede acertar o fallar. Ahora la frase que describe la situación sería:

A la vista de esta información Zaj no sabe dónde está el botín y decide entrar en la Norte.

Debemos distinguir estos tres componentes:

1. Dónde está realmente el botín. Es una realidad sobre la que Zaj no tuvo capacidad de decisión.
2. El grado de conocimiento o sospecha que Zaj tiene sobre esa realidad. El primer año los datos le llevan a suponer que está en la Norte, mientras que el segundo año los datos le mantiene en la incertidumbre.
3. La acción que Zaj decide ejecutar a la vista de la información que tiene o a pesar de no tener información suficiente.

Pronto veremos que en la elaboración de las conclusiones de los trabajos científicos entran en juego tres aspectos equivalentes a estos.

EPÍLOGO

Hemos visto que en la conversación coloquial el término «*decidir*» tiene un uso laxo y al construir las conclusiones de un trabajo esa palabra debe ser usada con más precisión para evitar incoherencias.

En el próximo capítulo reflexionamos detenidamente sobre las diferencias entre los «tests de significación», concebidos por Fisher para ayudarnos en nuestro intento de conocer mejor la Naturaleza y los «tests de hipótesis», diseñados por Neymann y E. Pearson para ayudarnos en las situaciones en que tenemos que tomar decisiones.

Aunque muchos estadísticos utilizan sistemáticamente el formato de los tests de hipótesis, asumiendo que las investigaciones científicas implican siempre una toma de decisiones en base a un valor de P que se convenga como frontera, ya sabemos que eso no es lo habitual en la investigación científica.

COMPRUEBE SU NIVEL DE CONOCIMIENTOS. ENCUESTA DE AUTOEVALUACIÓN

En el Apéndice 2 encontrará una encuesta de autoevaluación para este capítulo, que le ayudará a evaluar en qué medida tiene claras sus ideas en este tema.

Test de significación *versus* test de hipótesis

En el capítulo anterior reflexionamos sobre la diferencia entre *pensar* y *actuar*, es decir, entre hacer estudios para adquirir más conocimiento sobre un tema, propio de la investigación científica, y hacerlos para elegir una acción a ejecutar, propio de la toma de decisiones.

Ahora ya estamos en condiciones de comentar las diferencias entre **tests de significación** (Fisher, 1916) y **tests de hipótesis** (Neyman y E. Pearson, 1933).

Ello nos ayudará a deshacer un malentendido que durante decenios enturbió el proceso de la elaboración de conclusiones en la investigación científica al interpretarlo los investigadores como una cuestión de toma de decisiones. Entender las causas de los errores facilita decisivamente evitarlos en el futuro.

LA POLÉMICA. «TEST DE SIGNIFICACIÓN» Y «TEST DE HIPÓTESIS»

Entre 1916 y 1925 Ronald Fisher desarrolló los *tests de significación* (TS) para ayudar a elaborar conclusiones en la investigación científica. A él, a Student (J. Gosset) y a K. Pearson corresponde el mérito de encontrar las distribuciones estadísticas más frecuentemente implicadas en el cálculo del valor P de los tests.

«El fundamento lógico de los tests de significación», dice Fisher, «es el de rechazar hipótesis que solamente por coincidencia pudieran llevar a los datos observados»¹.

Por otra parte, en 1933 Neyman y E. Pearson publicaron su famoso trabajo proponiendo, con el nombre de *tests de hipótesis (TH)*, el uso del valor P del test para decantarse por una hipótesis o su alternativa, según sea mayor o menor que una determinada cantidad convenida al respecto, 0,01; 0,05 o cualquier otra.

Desde entonces se mantiene una tensa polémica entre valedores de uno y otro enfoque.

Veamos detalladamente estos dos puntos de vista y los argumentos a favor y en contra de cada uno, usando nuestro ejemplo de estudio de tres presuntos anticancerígenos (AC), cada uno de los cuales se prueba en 40 ratas de una cepa que desarrolla cáncer espontáneamente en el 60% de ellas. He aquí los resultados, el valor P del test y los intervalos de confianza al 95% y al 99%.

Fármaco	Núm. de ratas con cáncer en la muestra	% de ratas con cáncer en la muestra	Valor P	IC al 95%	IC al 99%
A	5	12,5%	0,000000003	4%-27%	3%-32%
B	18	45%	0,039	29%-62%	25%-66%
C	19	47,5%	0,074	32%-64%	27%-68%

A.1) Los partidarios de los TS subrayan que la evidencia contra la H_0 —que dice que el producto no es AC— *aumenta gradualmente* al disminuir el valor P, no habiendo un punto de corte que separe los valores P que llevan a rechazar la H_0 de los que invitan a aceptarla como posible. Enuncian las conclusiones en estos términos:

¹ Aunque el propio Fisher comentó que valores de P inferiores a 0,05 ya le parecían un notable argumento en contra de la hipótesis nula, nunca pretendió hacer de ese valor una barrera mítica, ni de esa reflexión un dogma de fe.

«Pensamos que “A” es un potente AC, mientras que “B” y “C” puede que sean AC o que no lo sean. Con este estudio es imposible formarse opinión al respecto».

- B.1) Los partidarios de los TH consideran que cuando el profesional debe tomar una decisión, necesita convenir un valor de P límite y rechazar la H_0 cuando P sea inferior a ese límite. Enuncian las conclusiones en estos términos:

«Habiendo decidido un valor alfa de 0,05, concluimos que el efecto AC observado en las muestras ha sido “estadísticamente significativo” para “A” y “B”, pero no para “C”».

- A.2) Los partidarios de los TS hacen estas objeciones: «¿Qué quiere decir “estadísticamente significativo”?». Indica simplemente que es $P < 0,05$ (o el límite que se haya convenido) y nuestra opinión sobre si un producto es o no es AC no se modifica sensiblemente porque P esté a uno u otro lado de 0,05. En vez de reducir la expresión del resultado a la dicotomía «significativo» o «no significativo», lo correcto es dar el valor P encontrado, que cuantifica el grado de evidencia contra la H_0 . En nuestro ejemplo los resultados no permiten tomar postura sobre si «B» y «C» son o no AC, y esa incertidumbre no se reduce por adjetivar el resultado como «significativo» o «no significativo».
- B.2) Pero los partidarios de los TH hacen estas objeciones: «Si los fisherianos dicen que valores P “muy pequeños” obligan a rechazar la H_0 , en algún punto debe empezar la zona de los valores “muy pequeños”. Es necesario acordar un valor P por debajo del cual se rechaza la H_0 ».

Vemos que aunque ambos enfoques utilizan el valor P del test, lo hacen de distinto modo y hay argumentos sólidos a favor y en contra de cada uno de ellos.

El verdadero damnificado de este pertinaz desencuentro entre los estadísticos partidarios de uno y otro enfoque es el investigador no versado en estadística, que se ve atraído por el procedimiento de los tests de hipótesis y lo emplea inadecuadamente en la elaboración de las conclusiones de sus trabajos. Al no entender los matices que entran en juego en esta polémica termina asumiendo la escueta simplificación que atribuye

«validez» a los resultados con $P < 0,05$ (u otro valor convenido) y la niega cuando P supera ese límite.

NO SON ENFOQUES CONTRAPUESTOS, SINO COMPLEMENTARIOS

Una reflexión cuidadosa nos muestra que ambas posturas no constituyen estrategias contrapuestas, sino complementarias, y ambas son correctas..., cada una en su contexto:

Test de significación → para **formarse opinión**, propio de la *adquisición de conocimiento* y

Test de hipótesis → para **elegir entre dos acciones**, propio de la *toma de decisiones*.

La adquisición de conocimiento y la toma de decisiones son dos situaciones diferentes que requieren diferentes estrategias de inferencia estadística, tanto en la vida común como en la investigación científica.

Para ver que ambos enfoques no son contrapuestos sino complementarios, recordemos que tanto en la vida común como en la investigación científica es muy frecuente que a partir de cierta información parcial sobre un tema el sujeto intente *formarse opinión* sobre aspectos no directamente conocidos y/o tomar una *decisión práctica* relacionada con ese tema. Distingamos entre estas dos finalidades:

Formarse opinión y Elegir una acción

En algunos casos la información obtenida nos lleva claramente a formarnos cierta opinión y ello permite elegir inequívocamente una determinada acción, si es el caso. Otras veces, la información no nos permite formarnos opinión y por ello no está claro qué acción debe elegirse, de modo que, si es posible, no se elige ninguna, y si es obligado elegir una, se hace sabiendo que hay cierto riesgo de equivocarse.

Volvamos al ejemplo de los tres posibles anticancerígenos (AC), teniendo ahora como objetivo que los productos que en ese estudio muestren resultados «prometedores» serán sometidos a otro estudio más detallado para precisar la magnitud de su efecto AC y los mecanismos de acción. Las dos finalidades posibles son:

Formarse opinión → ¿Es ese producto realmente AC o no lo es?

Elegir una acción → Pasar o no pasar el producto a estudio más detallado.

Veamos qué opinión nos formamos y qué decisión tomamos, teniendo en cuenta que espontáneamente hacen cáncer el 60% de las ratas.

Formarse opinión → *Test de significación*: es obvio que el resultado con «A» nos lleva a *pensar* que es AC, pues si no lo fuera sería extraordinariamente difícil que aparecieran por simple azar tan pocos cánceres como en esa muestra. Para «B» y «C» la situación no es clara. No es fácil, pero tampoco muy difícil, que aparezcan espontáneamente ese número de cánceres si los productos no fueran AC. El intervalo de confianza al 99% muestra que es posible que «B» incremente el % en 6 puntos o algo más, que no lo modifique o que lo baje en 35 puntos o algo más. Estas cantidades varían acordemente con la confianza elegida, pero para cualquiera que sea dicho nivel de confianza, los límites del intervalo no tienen carácter de valor frontera. Para «C» la situación es similar.

Elegir una acción → *Test de hipótesis*: el hecho de pensar que «A» es AC nos lleva a decidir sin titubeos que pase a la segunda fase, pero la incertidumbre que tenemos respecto a «B» y «C» no nos invita claramente a decidir que pasen o que no pasen. En estos casos hay riesgo de equivocarse en cada uno de estos dos sentidos:

1. Pasar a estudio posterior un producto que realmente no es AC y por ello no debería haber pasado (es el llamado error tipo I) y
2. Equivocarse por no pasar un producto que realmente es AC y por ello debería haber pasado (el llamado error tipo II).

¿Qué ayuda prestan los tests estadísticos en estas situaciones?: el valor P nos permite adoptar un criterio de acción que determina el riesgo de cometer error tipo I (pasar productos que realmente no son AC). Si, por ejemplo, el investigador decide pasar a estudio posterior los productos que den $P < 0,05$, a la larga pasarán solo un 5% de los productos que realmente no son AC. Y si decide pasar los productos que den $P < 0,08$, a la larga pasarán solo un 8% de los que realmente no son AC. A este valor de P que se acuerda como límite se le suele llamar «alfa».

¿Qué cantidad debe elegirse como punto de corte, alfa? No hay razones matemáticas ni estadísticas a favor de ninguna cantidad concreta. Lo

que la Estadística hace por el investigador es informarle del riesgo de equivocarse en función del criterio que elija. Es decisión del investigador establecer ese riesgo. En general, cuanto menor es α (para tener poco riesgo de pasar productos que realmente no son AC), mayor es el riesgo de cometer error tipo II, es decir, no pasar productos que realmente son AC².

Esta relación inversa —al disminuir un riesgo aumenta el otro— se presenta con frecuencia en la vida común. Por ejemplo, si un profesor pone muy alto el listón para aprobar, reduce el riesgo de que alumnos ignorantes aprueben por casualidad, pero aumenta el riesgo de que alumnos con conocimientos suficientes suspendan. Si baja el listón, aumenta el primer riesgo (aprobar inmerecidamente) y reduce el segundo (suspender inmerecidamente). ¿Y cuál es la nota más adecuada como barrera? Obviamente no hay razones objetivas a favor de ninguna cantidad concreta.

En la tabla siguiente resumimos los dos procedimientos en el ejemplo de los anticancerígenos.

	Informe parcial	Formarse opinión	Elegir una acción
	% Muestral, valor P, IC _{95%} e IC _{99%}	¿Pensamos que es AC?	¿Decidimos estudiar la sustancia detenidamente?
		T. de significación	T. de hipótesis
		No hay valor frontera, a menor P, más evidencia contra H ₀	Depende del valor frontera que se convenga (Alfa)
			Alfa 0,01 Alfa 0,05 Alfa 0,10
A	12,5% → P = 0,000000003 4%-27% y 3%-32%	Sí	Sí Sí Sí
B	45% → P = 0,039 29%-62% y 25%-66%	?	No Sí Sí
C	47,5% → P = 0,074 32%-64% y 27%-68%	?	No No Sí

² La probabilidad de cometer este tipo de error depende de cuánto ese producto disminuye realmente el % de cánceres.

RESUMIENDO: en la investigación científica hay que distinguir dos procesos distintos y complementarios: *adquisición de conocimiento y toma de decisiones fácticas*.

El investigador no versado en matemáticas puede entender con toda claridad la lógica propia de cada uno de estos procesos, que es sencilla y reproduce fielmente la usada continuamente en procesos equivalentes de la vida común. Pero a la mayoría de ellos no se les ha explicado con claridad esta diferencia y mezclan confusamente elementos de ambos procedimientos.

En el ejemplo de los anticancerígenos el hecho de que se decida, por ejemplo pasar a estudio posterior a los productos que en la prueba inicial dan $P < 0,05$ implica que «B» pase y «C» no pase. Pero ello no quiere decir que tengamos la seguridad de que «B» es AC y «C» no lo es. Respecto a ambos tenemos claras dudas y esas dudas no se eliminan por haber decidido que uno pase y otro no. También en el ejemplo del profesor es nítida esa diferencia. Él tiene claro que un alumno con nota 9,3 conoce la materia y uno con nota 1,4 no la conoce. Pero es consciente de que no hay una cantidad que marque el límite. Si finalmente decide poner la barrera en 5, el alumno con 4,9 suspenderá y el de 5,1 aprobará, pero ello no disminuye sus dudas acerca de esos dos alumnos, ni le proporciona la certeza de que el primero ignoraba la materia y el segundo la conocía. Una vez más, los puntos de corte son imprescindibles para decisiones fácticas, pero no desempeñan ningún papel en el proceso de formarnos opinión sobre un tema.

Recopilemos los distintos tipos de conclusiones que enunciarían diversos profesionales para el producto «B», que dio un valor $P = 0,039$ y el «C» que dio $P = 0,074$.

AUTOR	EXPRESIÓN	COMENTARIO
Estadístico aplicando un <i>Test de significación</i> encaminado a formarse opinión sobre si cada producto es o no AC	<i>«Imposible formarse opinión para «B» y «C» con estos resultados»</i>	CORRECTA
Estadístico aplicando un <i>Test de hipótesis</i> para decidir si cada producto pasa o no pasa a estudio más detallado	<i>«Habiendo convenido un valor alfa de 0,05, pasa «B», pero no pasa «C»</i>	CORRECTA
Investigador intentando elaborar conclusiones	1. <i>«Habiendo convenido alfa = 0,05, el efecto AC observado en las muestras es 'estadísticamente significativo' para «B», pero no para «C».</i>	ININTELIGIBLE, ¿Qué quiere decir «estadísticamente significativo»?
	2. <i>«Habiendo decidido considerar el resultado 'significativo' si es $P < 0,05$, concluimos que «B» es AC y «C» no es AC»</i>	INADMISIBLE Ambas afirmaciones son gratuitas, no avaladas por los datos

NOTA: dado que este punto es clave para entender cómo la Inferencia Estadística puede ayudar en la investigación científica y en la toma de decisiones, y la importancia de no confundir ambos procedimientos, insistimos en esta idea con un nuevo ejemplo que se incluye como apéndice al final de este capítulo. Está destinado al lector que necesite aclarar estas ideas. El que haya entendido nítidamente el apartado anterior no encontrará en este apéndice nada nuevo.

LO QUE DICEN ALGUNOS DE LOS MÁS CUALIFICADOS EXPERTOS

Ya vimos que las conclusiones de un trabajo científico se refieren al grado de seguridad que tenemos en que cierta hipótesis sea falsa y ello no es un tema de «decisión» sino de convicción³. Sin embargo, el enfoque Neyman-Pearson, adecuado para tomar decisiones, acabó siendo mayoritariamente usado para elaborar conclusiones de trabajos científicos que no implican toma de decisiones sino intento de conocer. Esta importación acrítica de una metodología desde el campo para la que fue creada a otro en el que no tiene validez ha tenido los efectos negativos que estamos comentando.

¿Y cuáles son las causas de que se originara, creciera y permaneciera la confusión entre estos dos procesos —formarse una opinión y decidir una acción— haciendo que muchos investigadores procedan como si de tomar decisiones se tratara, cuando en realidad se trata de formarse opinión sobre el comportamiento de la Naturaleza?

¿Cómo es posible que durante décadas la «regla del 5%» haya ofuscado la mente de los investigadores y amenace con seguir distorsionando por muchos años el proceso de la Inferencia Estadística? ¿Por qué los tests estadísticos, creados para ayudar a elaborar conclusiones más razonables y justificadas, son mal usados, llevando en muchos casos a conclusiones arbitrarias?

Fisher (1956) hace notar que Neymann y Pearson —además de no hacer aportaciones matemáticas que ayudaran al cálculo del valor P— no trabajaban con investigadores, es decir, no ayudaban a los científicos en el diseño de los estudios y en la generación de las conclusiones: *«Hacia 1930 todos los problemas estadísticos que merecían tratamiento cuidadoso ya habían sido discutidos en términos de tests de significación.... que constituían una eficaz ayuda en la interpretación de los datos. La teoría de los tests de hipótesis fue un intento posterior, a cargo de autores que no habían tomado parte en el desarrollo de los tests de significación ni en sus aplicaciones científicas, de reinterpretarlos para tomar decisiones,*

³ Veámoslo una vez más con un ejemplo de la vida común. La hipótesis que dicen que Juan tiene 50 años será rechazada si nos dicen que es el actual campeón olímpico de velocidad. Pero no se trata de que nosotros *decidamos* la edad de Juan. Lo que hacemos es *pensar* que tiene menos de 50 años, porque el dato que observamos es incompatible con que los tenga.

aunque el proceso lógico de la toma de decisiones es muy diferente del que usa el científico que intenta conocer mejor la realidad».

Si el investigador le pide a la Estadística un procedimiento para tomar decisiones, esta debe ofrecerle el más eficiente, que es el propuesto por Neymann y Pearson con el nombre de test de hipótesis. Es la herramienta más adecuada para ayudar a elegir una acción, cuando hay que hacerlo.

Pero no se puede confundir la mecánica operativa de la toma de decisiones (se decide una acción u otra según el valor P esté a uno u otro lado de una cifra frontera) con el proceso mental de la elaboración de conclusiones razonables, en el que no se trata de decidir nada, sino de ir conociendo mejor la naturaleza.

Snedecor y Cochran (1950), dos de los estadísticos más relevantes del siglo xx dicen que *«debe evitarse esa actitud que consiste en considerar los tests de significación como una regla para decidir de modo automático si se acepta o se rechaza una hipótesis»*.

Más contundentes son las palabras de K. Rothmann (1986), una de las mentes más lúcidas en el análisis de datos biomédicos, cuando dice: *«Algunos problemas de la industria y la agricultura implicaban experimentos en base a los cuales había que elegir entre dos posibles acciones. Los experimentos fueron diseñados para producir resultados que permitieran tomar decisiones y los conceptos heredados de estos orígenes aún están presentes en la investigación científica actual»*.

Podría pensarse, por ejemplo, que los ensayos clínicos (EC) se hacen para *«decidir»* si en el futuro se debe usar uno u otro tratamiento. Y eso es cierto como idea general que abarca a un conjunto de EC sobre el mismo tema, pero de *cada EC en particular* no se espera que lleve a tomar una decisión en ese sentido. En palabras de Rothman (1998):

«Cuando un solo estudio constituye el único elemento para decidir entre dos posibles acciones, como en el control de calidad en la industria, el enfoque de toma de decisiones del análisis estadístico está justificado. Pero en la investigación científica el razonamiento correcto requiere más que la clasificación en “significativo” o “no significativo”. La degradación de la información en una simple dicotomía es contraproducente... Es presuntuoso, sino absurdo, por parte del investigador pensar que los resultados de su estudio serán la única base para tomar

decisiones científicas. Tales decisiones se toman en base a los resultados de muchos estudios».

Insistamos en esta idea que es clave para evaluar el tipo de asistencia que la Inferencia Estadística puede y debe prestar a la investigación científica: la inmensa mayoría de las investigaciones en el campo de las ciencias de la salud (y de las demás ciencias) no se hacen para *tomar decisiones inmediatas* basadas en el resultado de un estudio, sino para aportar información al *conocimiento* de un tema. Silva y Rozeembon (1997) subrayan: *«La tarea del científico no es prescribir acciones, sino establecer convicciones razonables. El propósito central de un experimento no es precipitar la toma de decisiones, sino propiciar un reajuste en el grado de confianza que uno tiene en la veracidad de cierta hipótesis... y la creencia en una proposición no es un asunto de todo o nada».*

LOS INVESTIGADORES RENUNCIAN A RAZONAR

Otra causa del arrollador éxito que la «regla del 5%» tuvo entre los investigadores de todas las disciplinas radica en que el procedimiento operativo de la toma de decisiones es muy atractivo por su simplicidad y porque mediante él toda investigación puede ser publicada con conclusiones *aparentemente* claras y rotundas.

Aferrándose a esa receta los investigadores pueden obviar las limitaciones propias de los tests de significación —que con un valor de P grande o intermedio no permiten pronunciarse definitivamente ni a favor ni en contra de la hipótesis— y consiguen fabricar conclusiones presuntamente claras en todos sus trabajos.

Una vez más Rothman (1986) pone el dedo en la llaga cuando comenta: *«¿Por qué esa dicotomización —resultado significativo o no significativo— se ha hecho tan popular en la investigación científica? Evidentemente en gran parte por la aparente objetividad y nitidez que implican esas expresiones. Sustituyen la reflexión razonable acerca de los resultados por la aplicación mecánica de unas palabras. Editores, investigadores y lectores prefieren la aparente rotundidad de esas expresiones a una valoración realista que no permite encasillar los resultados en buenos o malos».*

Esta actitud de los investigadores viene favorecida por el temor a que usar el valor P razonablemente requiera conocimientos matemáticos que

ellos no tienen. Ante la imposibilidad de entender el lenguaje algebraico con que se les intenta explicar esta materia, no tienen otra salida que refugiarse en la repetición mecánicamente de una rutina adoptada como dogma de fe. Asumen erróneamente que razones de gran envergadura matemática avalan procedimientos como la «regla del 5%», dado que los estadísticos la usan y no son conscientes de que los estadísticos las usan con un enfoque muy distinto.

UN RESULTADO CON VALORES P GRANDE NO PERMITEN TOMAR POSTURA PERO PUEDEN CONSTITUIR INFORMACIÓN MUY ÚTIL

Ciertamente, no es fácil privar al investigador de esa muleta en la que lleva apoyado diez, veinte o treinta años de vida profesional. Al quedarse sin la sencilla— y falsa— «regla del 5%» para elaborar conclusiones, se siente huérfano e inerte.

Además, teme que los estudios con valores de P no muy pequeños carezcan de interés. Si con valores de P grandes o intermedios —que son habituales en la investigación real— no es posible decantarse ni a favor ni en contra de una hipótesis, parece que carecen de valor la mayoría de los estudios. ¿Deben ser desechados como inservibles los experimentos en los que P no es extremadamente pequeña? ¿Qué podemos publicar cuando el valor P del test es mayor de 0,05, lo cual ocurre en muchos estudios?

Debe estar claro que los estudios que no permiten rechazar la hipótesis nula porque el valor P no es extremadamente pequeño contienen información que puede ser muy útil. Considerado aisladamente, cada estudio no permite llegar a una conclusión contundente, pero considerado *junto con otros trabajos* sobre el mismo tema puede ser decisivo para aclarar el problema investigado. La comunidad científica rara vez se deja convencer por los resultados de un solo estudio. Por el contrario, tiene en cuenta *los resultados de estudios similares* y toma postura a favor de que cierto efecto es una realidad general cuando se encuentra ese efecto reiteradamente en varios estudios. Se asume que en la mayoría de los casos los resultados de cada trabajo individualmente considerado no van a permitir conclusiones definitivas.

Lo sensato es dar el valor de P encontrado. De ese modo el lector maduro puede valorar por sí mismo cuanta evidencia constituyen esos datos contra la hipótesis nula, no viéndose obligados, el autor ni el lector, a decantarse por una u otra hipótesis cuándo los datos disponibles no permiten hacerlo. Además, siempre que los datos lo permitan, deben darse y comentarse los intervalos de confianza, que constituyen una ayuda decisiva en la inferencia.

BIBLIOGRAFÍA

- Fisher RA. *Statistical methods for research workers*. Hafner Press, 1925.
— *The design of experiments*. Hafner Press, 1935.
— *Statistical methods and scientific inference*. Hafner Press, 1956.
Neyman J, Pearson E. «On the problem of the most efficient tests of statistical hypothesis». *Philosophical trans of the Royal Society of London A*. 1933 231: 289-337.
Rothman K and Greenland W. *Modern Epidemiology*. Lippincott-Raven Pub., 1998.
Rothman K. *Modern Epidemiology*. Little Brown. Toronto, 1986.
Rozeboom WW. «The fallacy of the null hypothesis significance test». *Psychological bulletin* 1960 56: 26-47.
Silva, L C *Cultura estadística e investigación científica*. Díaz de Santos, 1997.
Snedecor G y Cochran WG. *Statistical Methods*. John Wiley and Sons, 1950.

APÉNDICE: EJEMPLO DE TEST DE SIGNIFICACIÓN Y TEST DE HIPÓTESIS

Distintos proveedores ofrecen al hospital de Río Hacha, a muy bajo precio, varias sacas con cien mil agujas cada una, asegurando cada proveedor que su saca no contiene más de 10% de agujas defectuosas (AD). A Ana Buendía, estadística del hospital, se le pide que tome una muestra aleatoria de $N = 50$ agujas de cada saca, y de acuerdo a lo observado en ella haga dos cosas:

- Emitir juicio sobre si esa saca tiene o no más del 10% de AD.
- Decidir si se compra o se rechaza cada saca.

Llamaremos sacas «correctas» a las que realmente tienen un 10% de AD. Para cada saca Ana plantea la H_0 : «La saca es correcta, es decir, tiene 10% de AD» y calcula el valor P del test correspondiente.

Estos son los resultados del estudio de 4 sacas:

La H_0 es que cada saca tiene 10% de AD, no más. Se toman 50 agujas de cada saca

Saca	Num. de AD	% de AD	Valor P	IC al 95%	IC al 99%
A	49	98%	3×10^{-11}	89%-99%	86%-99%
B	10	20%	0,025	10%-34%	8%-38%
C	9	18%	0,060	9%-31%	7%-36%
D	6	12%	0,380	5%-24%	3%-28%

Veamos los dos enfoques:

- a) *Tests de significación* para intentar saber si hay o no más de 10% de AD:

Ana piensa que «A» contiene más del 10% de AD, mientras que para «B, C» y «D» no puede formarse opinión sobre si tienen o no más del 10% de AD. El valor P y los IC indican que el % de AD en cada una de esas sacas puede ser menor, igual o mayor que el 10%.

- b) *Tests de hipótesis* para decidir si se rechaza o se compra cada saca:

Ana decidirá rechazar «A», pero para «B, C» y «D» la información obtenida no lleva a una postura clara.

En estos casos hay riesgo de equivocarse por rechazar sacas correctas (error tipo I) o equivocarse por comprar sacas no correctas (error tipo II).

La ayuda que prestan los tests estadísticos en esta situación es permitirnos adoptar un criterio de acción que determina el riesgo de cometer error tipo I (rechazar sacas correctas). Si, por ejemplo, el investigador decide rechazar las sacas que den $P < 0,07$, a la larga rechazará solo un 7% de las correctas. Y si decide rechazar las sacas que den $P < 0,009$, a la larga rechazará solo 9 de cada mil correctas.

Recuerde que no hay razones matemáticas ni estadísticas a favor de ninguna cantidad concreta como punto de corte. Lo que la Estadística

hace por el investigador es informarle del riesgo de equivocarse en función del criterio que elija. Es decisión del investigador establecer ese riesgo. En general, cuanto menor es alfa (para tener poco riesgo de rechazar sacas correctas), mayor es el riesgo de cometer error tipo II, es decir, comprar sacas que tienen más del 10% de AD.

Resumamos lo que Ana puede *pensar* y *hacer* con cada saca:

				Test de significación ¿Pensamos que hay más de 10% de AD en la saca?	Test de hipótesis ¿Decidimos rechazar la saca?
				No hay valor frontera	Depende del valor frontera que se convenga
				A menor P, más evidencia contra H ₀	Alfa 0,01 Alfa 0,05 Alfa 0,10
% de AD	Valor P	IC al 95%			
A 98%	3 × 10⁻¹¹	89-99		Sí	Sí Sí Sí
B 20%	0,025	10-34		?	No Sí Sí
C 18%	0,060	9-31		?	No No Sí
D 12%	0,380	5-24		?	No No No

De nuevo, el error en el que caen muchos investigadores es confundir la barrera que se pueda poner para decidir la compra o rechazo de cada saca con el grado de certeza o duda que tengamos al respecto. Si, por ejemplo, se decide rechazar las sacas en las que aparezca $P < 0,10$, solo se compraría la «D», pero ello no implica que «B» y «C» sean incorrectas. Pueden serlo y pueden no serlo. Y la «D» puede también ser incorrecta. Recuerde los intervalos de confianza para el % real de AD en «D». El punto de corte que elijamos no afecta en modo alguno a nuestras certezas e incertidumbres. Solo vale para determinar las sacas que compramos y las que rechazamos, pero somos conscientes de que podemos estar rechazando algunas sacas correctas y comprando algunas incorrectas.

Es válido decir, por ejemplo, «Habiendo convenido un valor alfa de 0,10, se compra «D» y se rechazan «A, B» y «C», pero es totalmente inadecuado decir «Habiendo convenido un valor alfa de 0,10, concluimos que «D» es correcta y no lo son «A, B» y «C».

Recopilamos en la siguiente tabla los distintos tipos de conclusiones que enunciarían diversos profesionales para la saca «B», que dio un valor $P = 0,025$ y la «C» que dio $P = 0,060$ y «D» que dio $P = 0,380$.

AUTOR	EXPRESIÓN	COMENTARIO
Estadístico aplicando un <i>Test de significación</i> encaminado a formarse opinión sobre cada saca	<i>«Imposible formarse opinión sobre “B, C” y “D”»</i>	CORRECTA
Estadístico aplicando un <i>Test de hipótesis</i> para decidir si cada saca es o no comprada	<i>«Habiendo convenido un valor alfa de 0,05, se compran “C” y “D” y no se compra “B”»</i>	CORRECTA
Investigador intentando elaborar conclusiones	1. <i>«Habiendo convenido alfa = 0,05, el exceso de AD observado en las muestras es ‘estadísticamente significativo’ para “B”, pero no para C».</i>	ININTELIGIBLE, ¿Qué quiere decir «estadísticamente significativo»?
	2. <i>«Habiendo decidido considerar el resultado ‘significativo’ si es $P < 0,05$, concluimos que “C” y “D” son correctas y “B” no lo es»</i>	INADMISIBLE

Lo que no es el valor P del test

Ahora ya sabemos que el valor P del test, usado para realizar los tests de significación (TS) en investigación y elaborar las conclusiones, nos dice la probabilidad de obtener cierto tipo de muestras cuando es cierta la H_0 .

Un error muy frecuente es creer que el valor P del test indica la probabilidad de que la hipótesis nula (H_0) sea cierta o sea falsa.

Este capítulo lo dedicamos a ver la diferencia entre esas dos probabilidades:

1. La probabilidad de que la hipótesis nula sea cierta.
2. La P del test, que es la probabilidad de que aparezca cierto tipo de muestras cuando es cierta la hipótesis nula.

Comenzaremos a aclarar esa confusión en los siguientes ejemplos de la vida común.

PROBABILIDAD DE UN SUCESO Y PROBABILIDAD DE UN SUCESO CONDICIONADO A OTRO

Ejemplo 1.º El señor «A» nos dice:

«Si mañana llueve la probabilidad de que haya un accidente de tráfico en la calle Pez es $P = 0,57$ ».

¿Nos ha hablado sobre la probabilidad de que llueva? Ciertamente no. Nada se ha mencionado sobre eso. ¿Qué evento tiene una probabili-

dad de 0,57? Respuesta: 0,57 es la probabilidad de que haya un accidente de tráfico en esa calle *si llueve*.

Ejemplo 2.º El señor «B» nos dice:

«He comprado lotería y si me toca es muy probable que me vaya al Caribe».

¿Nos ha hablado sobre la probabilidad de que le toque la lotería? Por supuesto que no. No nos dijo si había comprado muchos o pocos décimos y por ello no sabemos nada acerca de la probabilidad de que le toque o no le toque la lotería. ¿Qué cosa es muy probable? Respuesta: que vaya al caribe *si le toca la lotería*.

Ejemplo 3.º El niño de la señora «C» tiene un ligero dolor abdominal. Como hace unos días ella vio que el niño de su vecina tuvo apendicitis aguda (AA) le pregunta al doctor si cree que su niño puede tener AA y qué habría que hacer si la tuviera. El doctor responde que la gran mayoría de dolores abdominales infantiles son episodios banales que remiten espontáneamente. Y añade:

«Pero si el niño tiene AA es muy probable que sea operado».

¿Dijo el médico que es muy probable que el niño tenga AA? Respuesta: ¡NO! ¿Qué cosa es *poco* probable? Respuesta: que el niño tenga AA ¿Qué cosa es *muy* probable? Respuesta: que haya que operar *si tiene AA*.

En los tres ejemplos anteriores, y en general, nunca se puede confundir estas dos cosas:

1. Probabilidad de que ocurra cierto suceso, «H».
2. Probabilidad de que ocurra el suceso K, *si ocurre* «H».

Insistamos en esta idea considerando el caso en que en cierta enfermedad se curan el **20%** de los pacientes no tratados. Para ver si el fármaco «A» incrementa el % de curaciones, se le administra a 5 enfermos y se encuentra que se curan los 5.

Planteamos la hipótesis nula: *«El tratamiento «A» es inútil, es decir, con él curan el 20% de los enfermos».*

Si esto fuera cierto, entre 5 enfermos esperamos, teóricamente, encontrar *una* curación (1 es el 20% de 5). Y un pequeño cálculo nos dice que si con «A» se curan el 20% y lo probamos en muchos grupos de

5 enfermos, en solo 3 cada 10.000 de esos grupos aparecerán curados los cinco. Es decir, valor P del test: $P = 0,0003$.

¿A qué cosa se refiere esa probabilidad de $P = 0,0003$?

- a) Que el medicamento «A» sea inútil (que con «A» se curen 20%).
- b) Obtener curación en los 5 enfermos si el medicamento «A» es útil.
- c) Que el medicamento «A» sea útil (que con «A» se curen más de 20%).
- d) Obtener curación en los 5 enfermos si el medicamento «A» es inútil¹.

EL VALOR P Y LA PROBABILIDAD DE QUE LA HIPÓTESIS NULA SEA CIERTA O SEA FALSA

En la inmensa mayoría de los estudios biomédicos no se puede calcular la probabilidad de que la hipótesis planteada sea cierta. Veremos un ejemplo muy sencillo de otro área en el que sí se puede calcular la probabilidad de que la H_0 sea cierta. Ello nos ayudará a distinguir esa probabilidad de la P del test y nos ayudará a entender por qué utilizamos esta última para elaborar conclusiones razonables.

Sabemos que el 3% de las monedas de una ciudad son falsas, y en ellas la probabilidad de salir «cara» es mayor de 0,5 .

- 1.º Hemos comprado una moneda en la ciudad y no sabemos si es legal o falsa. Consideremos la hipótesis nula, H_0 , que dice:

La moneda es legal, es decir, la probabilidad de salir cara es 0,5.

¿Cuál es la probabilidad de que H_0 sea cierta? Si son falsas el 3% de las monedas, son legales el 97%. Por tanto, la probabilidad de que la H_0 sea cierta, es decir, que la moneda sea legal es del 97%².

- 2.º Para intentar saber si es falsa o legal hacemos 14 tiradas con ella y encontramos que en las 14 tiradas sale «cara». ¿Nos invita este resultados a pensar que la moneda *no es legal* (rechazar H_0), o más bien sugiere que la moneda *puede ser legal* (aceptar H_0 como posible)?

¹ La correcta es la d.

² Esa probabilidad $P = 0,97$, quiere decir que de cien personas que compren una moneda en esa ciudad, 97 tendrán una moneda legal.

La P del test es $P = 0,00006$. Es decir, haciendo millones de series de 14 tiradas cada una con una moneda legal, solamente en 6 cada 100.000 series aparecen cara en los 14 lanzamientos. Lo cual es una notable evidencia en contra de la H_0 , es decir, a favor de que la moneda es falsa.

Distinga claramente entre:

- $P = 0,97 \rightarrow$ es la probabilidad de que la H_0 sea cierta, es decir, que la moneda sea legal.
- $P = 0,00006 \rightarrow$ es la probabilidad de obtener todo caras, que es lo que ocurrió con nuestra moneda, *si la moneda es legal*.

La probabilidad de que la moneda sea legal es $P = 0,97$, lo cual indica que de cada 100 personas que compren una moneda en esa ciudad, 97 la tendrán legal, y nos proporciona bastante confianza en que nuestra moneda sea legal, pero no nos asegura que lo sea.

Para intentar saberlo hacemos el experimento con *nuestra moneda* y nos inclinaremos a rechazar la hipótesis de que es legal porque si lo fuera sería muy difícil que saliera cara en los 14 lanzamientos.

Fíjese en que si no conociéramos la proporción de monedas legales, es decir, la probabilidad inicial de que la moneda comprada por nosotros fuera legal, nuestra conclusión sería la misma.

En la investigación científica se procede y razona de un modo muy semejante a este ejemplo. La mayoría de las veces no se puede evaluar la probabilidad de que la hipótesis nula sea cierta o sea falsa (equivale a no tener información sobre la proporción de monedas falsas en esa ciudad). El investigador toma una muestra, lo que equivale a tirar la moneda 14 veces, y calcula la probabilidad de que aparezca por azar el tipo de resultado encontrado por él, si es cierta la hipótesis nula. Rechaza la H_0 cuando esa probabilidad es muy pequeña. Y si es grande dice que la H_0 puede ser cierta.

UN EJEMPLO MÉDICO CON PROBABILIDAD DE LA HIPÓTESIS

Se sabe que el 4% de los habitantes de una ciudad portan la variedad genética «HH-3», que hace que la Gammaglobulina G4 esté aumentada

en suero y favorece la aparición de demencia senil precoz (DS). La población con el gen normal tiene media de $G4 = 200$ y desviación estándar = 20.

Queremos ver si el señor «A» tiene esa variedad genética. Estudiar el gen es muy costoso e intentamos saberlo a través de su nivel de $G4$ en suero. Plantemos la hipótesis nula que dice que «A» *no tiene HH-3*.

1. La probabilidad de que esta H_0 sea cierta es $P = 0,96$, es decir, si tomamos 100 personas al azar, 96 de ellas no padecen HH-3. En principio, es bastante probable que A no tenga HH-3, pues solo 4 de cada 100 lo tienen.
2. Medimos su nivel de $G4$ en suero y encontramos $G4 = 300$. Un cálculo muy sencillo —cuyos detalles no hacen al caso ahora— nos dice que el valor P del test es $P = 0,000002$, es decir, 2 por millón.

Note la diferencia entre los dos valores de P que entran en juego:

1. $P = 0,96$ es la probabilidad de que «A» sea *normal*.
2. $P = 0,000002$ es la probabilidad de que un señor normal, que no tienen HH-3, tenga $G4$ por encima de 300.

Antes de medir la $G4$ considerábamos muy probable que «A» fuera normal. Pero a la vista de que su $G4$ es 300, nos inclinamos decididamente a pensar que tiene HH-3.

Sería un enorme error conceptual pensar que $P = 0,000002$ es la probabilidad de que H_0 sea cierta, pero, en este ejemplo, si el investigador cayera en él, no afectaría seriamente a la conclusión, ya que si la probabilidad de que H_0 sea cierta fuera tan pequeña lo lógico sería rechazarla, y esto es precisamente lo que se hace cuando se interpreta ese valor P correctamente.

Por el contrario, cuando el valor P del test es muy grande, confundirlo con la probabilidad de que H_0 sea cierta puede llevar a conclusiones equivocadas. Veamos un ejemplo con valor P grande. Para ello consideremos una situación semejante a la anterior pero con otra variación del gen, llamada HH-5 que, además de DS, produce *modificación* de la concentración de $G4$, aumentándola en unas personas y disminuyéndola en otras. En este caso no se conoce el % de casos con el defecto y no sabemos cuál es la probabilidad de que tenga HH-5 una persona toma-

da al azar. Estudiamos al Sr. «B» y encontramos que tiene $G4 = 200,1$. El test nos da $P_{BILATERAL} = 0,996$, (99,6% o 996 por mil). Creer que esa es la probabilidad de que la H_0 sea cierta, es decir, de que «B» sea normal, nos llevaría a tener una gran confianza en que lo fuera, lo cual es erróneo.

Esa probabilidad nos dice que de cada 1.000 personas normales 996 tienen $G4$ tan alejado de 200 como lo tiene «B» o aún más alejado, es decir, que 996 de cada 1.000 personas normales tienen $G4$ mayor de 200,1 o menor de 199,9. Lo cual nos lleva a pensar que es perfectamente posible que «B» sea normal, que no hay ninguna evidencia en contra de ello. Pero eso no equivale a decir que los datos son una fuerte evidencia a favor de que «B» sea normal³.

El intervalo de confianza al 99% para el valor medio de $G4$ en el colectivo al que «B» pertenece (no sabemos si el colectivo de normales o el de los que tienen HH-5) es: 150 a 250. Por tanto, podría ser que ese colectivo tuviera media de, por ejemplo, 235, lo que representa 35 unidades más que en los normales y podría responder a la presencia de HH-5. O podría ser que ese colectivo tuviera media de 160, es decir, 40 unidades menos que en los normales. Y, por supuesto, podría ser que ese colectivo tuviera media de 200, es decir, que fueran normales.

LA DIFICULTAD DE CONOCER LA PROBABILIDAD DE QUE SEA CIERTA LA H_0

Este tipo de ejemplos son una de las pocas situaciones en las que cabe calcular la probabilidad de que la hipótesis nula sea cierta. Recuerde que en la práctica la «probabilidad» es una frecuencia relativa, que se puede expresar como proporción (tanto por uno) o porcentaje (tanto por cien)... Pero en todo caso, para hablar de probabilidad tiene que estar definido de qué colectivo de «entes» se trata y qué les ocurre a una parte de ellos.

En la mayoría de las situaciones de la investigación científica no cabe hablar de la probabilidad de que la hipótesis nula sea cierta. Por ejemplo,

³ La radical diferencia que hay entre «no hay nada en contra esa hipótesis» y «es muy probable que esa hipótesis sea cierta» fue comentada con detalle en el Capítulo 9. Si el lector tiene dudas al respecto puede aclararlas leyendo ese capítulo detenidamente.

si de un nuevo fármaco se supone que puede ser más efectivo que el tratamiento clásico para curar cierta enfermedad y se hace un estudio comparando ambos productos, la H_0 es que *ambos son igual de eficaces*. ¿Cómo se puede evaluar la probabilidad de que esa hipótesis es cierta? Si alguien nos dijera que esa probabilidad es, por ejemplo, $P = 0,17$, es decir, 17 cada 100. ¿A qué cien «entes» se refiere esa cifra y qué les ocurre a 17 de ellos? No parece haber respuesta para tan básica pregunta. El problema no es, pues, que sea difícil calcular esa probabilidad. El problema es que no cabe pensar en ella. Una frase como «de 100 fármacos como este, 17 son mejor que el clásico» no parece tener sentido.

En resumen: en la gran mayoría de los casos no conocemos, ni siquiera aproximadamente, la probabilidad de que la H_0 sea cierta. La estrategia del investigador es tomar una muestra y concluir que la H_0 es falsa si la muestra es incompatible o muy difícilmente compatible con ella. El valor P del test nos informa acerca de esa compatibilidad o incompatibilidad entre la H_0 y la muestra, dándonos la probabilidad de obtener muestras como esa o aún más alejadas de lo esperado bajo la H_0 cuando esta es cierta.

COMPRUEBE SU NIVEL DE CONOCIMIENTOS. ENCUESTA DE AUTOEVALUACIÓN

En el Apéndice 2 encontrará una encuesta de autoevaluación para este capítulo, que le ayudará a evaluar en qué medida tiene claras sus ideas en este tema.

El enigma del tamaño de la muestra

NOTA PREVIA: Pocos temas en el ámbito de la investigación han sido y son tan mal conocidos y erróneamente interpretados como este del tamaño de muestra mínimo necesario para que el estudio realizado permita extraer «conclusiones válidas».

Todo investigador que se dispone a hacer un trabajo debe decidir cuántos individuos va a estudiar y se pregunta cuál es la cantidad «adecuada». Es una pregunta lógica e incluso obligada.

Y aquí, como en el caso de la interpretación del valor P de los tests, el investigador tiende a buscar una respuesta dicotómica esquemática, creyendo que hay un valor de tamaño idóneo que proporciona máxima información a mínimo costo. Pero la realidad no es tan simple y la respuesta correcta suele desconcertarle y dejarle insatisfecho. Mucho más satisfecho se queda cuando se le da un tamaño que, presuntamente, es el «adecuado».

En este capítulo explicamos cuál es la realidad al respecto y por qué en la mayoría de los casos el tamaño de muestra no debe determinarse de acuerdo a una fórmula estadística, sino a otros criterios. Como en el resto del libro, no se explican fórmulas concretas, sino que se discuten los conceptos básicos implicados y el uso correcto de los resultados obtenidos. La experiencia docente muestra que para el investigador es muy difícil prescindir de ciertos prejuicios erróneos y aceptar un enfoque más realista y menos esquemático. Para intentar vencer esta resistencia hemos optado en este último capítulo por un lenguaje informal y la exposición de las ideas básicas a través de ejemplos pintorescos que hagan

más amena la lectura y permitan enfatizar los puntos clave. La aparente trivialidad de los supuestos considerados no debe en ningún momento hacer pensar al lector que el tema es de importancia menor. De hecho, es una preocupación constante de todos los investigadores, y además los comités científicos suelen equivocarse manifiestamente cuando exigen que el tamaño de la muestra elegido en un proyecto de investigación sea justificado con «criterios estadísticos rigurosos» o con «criterios científicos». De modo que mientras estos comités preguntan al investigador por qué ha decidido cierto tamaño de muestra, los estadísticos se preguntan: ¿por qué los miembros de las comisiones evaluadoras de proyectos tienen ideas tan erróneas respecto a los tamaños de muestra?

EL TERRIBLE DILEMA DE AÍDA BUENDÍA

Este supuesto, más realista que mágico, puede presentarse en cualquier proyecto de investigación. Jefa del Servicio de Farmacia del Hospital de Macondo, Aída decide hacer un estudio para estimar con precisión la proporción de personas que consumen hipnóticos (CH) en la población constituida por todos los habitantes del Caribe.

Para determinar el número de personas que debe encuestar consulta a un famoso bioestadístico, que tras comentar con ella el caso detenidamente le dice que debe entrevistar a 27.225 personas. Pero en el camino de vuelta a casa coincide en el tren con otro afamado bioestadístico, que tras comentar con ella el caso detenidamente, le dice que debe entrevistar a 50 personas.

«Sin duda —pensó nuestra heroína— cometí algún error al dar los datos que me pedía alguno de los expertos consultados. O quizás alguno se despistó al hacer sus cálculos». Así que volvió a hablar con cada uno de ellos. Pero, para su sorpresa, ambos se ratificaron en su cantidad.

Convocado un comité de los mejores expertos europeos para dirimir estas diferencias, su informe dijo que ambos estadísticos habían actuado correctamente.

Aída no da crédito a lo que sus ojos ven cuando lee esta sentencia. Es imposible que dos cantidades tan distintas puedan ser ambas correctas. Quizá una epidemia de insensatez asola a nuestros científicos, en aquesta hora de cierta confusión general. Pero la confusión de Aída se convierte en estupor y perplejidad cuando descubre que el informe termina sugiriendo que un tamaño adecuado podría ser $N = 374$.

Dado que, al parecer, la vieja Europa padece demencia senil generalizada, Aída convence a las autoridades para convocar a los mejores expertos del mundo. De EE UU, Canadá, India y Japón llegan los más autorizados profesionales para dirimir cuál es el tamaño de muestra correcto. Al fin las cosas quedarán aclaradas, piensa Aída, y los ineptos que tan flagrantes tonterías dijeron serán puestos en evidencia, descalificados y desposeídos de sus titulaciones académicas.

El Comité Internacional se reúne a las 9:00 horas. Y a las 9:05 horas, tras 5 minutos dedicados a leer los informes anteriores sobre el tamaño de la muestra, emite veredicto diciendo que todos ellos son correctos. El primer bioestadístico consultado ($N = 27.225$), el consultado en el tren ($N = 50$) y el comité de expertos europeos ($N = 374$), todos actuaron correctamente. Y en una nota adjunta este Comité Internacional dice que un tamaño adecuado podría ser $N = 1.040$.

Aída no sale de su asombro, perplejidad y enojo ¿Qué se puede pensar de esos —presuntos— científicos que tan groseras contradicciones dicen y sostienen? A la vista de que la locura y total enajenación parecen ser generalizadas Aída renuncia al cálculo teórico del «tamaño adecuado» de la muestra y decide actuar de acuerdo con los recursos humanos y económicos de que dispone.

Encuesta a una muestra aleatoria de $N = 1.000$ personas y encuentra que 200 de ellas son CH. Y publica que en la muestra estudiada fueron CH el 20% y el intervalo de confianza para el % poblacional fue: $IC_{95\%} (\%_{POBLAC}) = 17,5\% \text{ y } 22,5\%$.

Pero el Ministerio de Sanidad dice que ese dato carece de valor, puesto que el tamaño de la muestra fue decidido arbitrariamente por Aída, sin «criterios científicos» adecuados.

EL CORONEL NO TIENE QUIEN LE DÉ UN TAMAÑO DE MUESTRA

Por esa misma época el coronel médico Gerineldo Márquez, sabe que el tratamiento «A» evita en torno al 30% de los casos de carcinoma inducido en ratas por altas dosis de radiación Beta y cree que el tratamiento «B» puede evitar mayor porcentaje de cánceres que el «A». Para comparar la eficacia de ambos tratamientos decide hacer un experimento usando $N = 11$ ratas en cada grupo, pues el asesor estadístico

de su departamento estima que de ese modo tiene una potencia estadística del 85%¹.

Pero el Comité de Investigación de la Facultad de Medicina dice que para tener potencia del 85% debe incluir $N = 83$ ratas en cada grupo.

Márquez apela al comité científico de la Universidad y le responden que para tener potencia de 85% debe incluir $N = 265$ ratas en cada grupo.

Gerineldo sospecha que un Alzheimer precoz y galopante asola a los bioestadísticos del país y decide recurrir a sus contactos en Nueva Inglaterra. Al grito de «Huston, Huston, tenemos un problema», convoca a los mejores especialistas de EE UU y el UK. También la reunión de este consejo de sabios duró 5 minutos. Y contestaron que todos los tamaños de muestra sugeridos anteriormente han sido correctamente calculados y por tanto pueden considerarse válidos. El informe del comité termina con una nota diciendo que para tener una potencia del 85% un buen tamaño de muestra sería $N = 1.357$.

A partir de este momento se pierde el rastro de Gerineldo. Algunos datos sugieren que profesó como monje trapense, otros que se hizo el harakiri y no faltan indicios que le sitúan cantando tangos por la voluntad en un burdel anónimo de Montmatre.

LA SOLUCIÓN DEL ENIGMA

En las dos situaciones anteriores llama poderosamente la atención que al preguntar un investigador por el tamaño adecuado que debería tener su muestra para ser «suficientemente representativa», se le contesta con varias cantidades muy diferentes y asegurándosele que *todas son correctas*. Esa respuesta desconcierta a cualquier investigador y la experiencia docente demuestra que la mayoría de ellos tienen notable dificultad para entender la explicación de esas aparentes contradicciones. De hecho, una alta proporción de profesionales hacen oídos sordos a estos argumentos e insisten en permanecer asumiendo una idea radicalmente equivocada que la caprichosa diosa Fortuna ha propagado por doquier en la segunda mitad del pasado siglo.

¹ La potencia es la probabilidad de obtener diferencias «estadísticamente significativas», es decir, un valor P del test menor que cierta cantidad convenida, si realmente hay cierta diferencia entre ambos tratamientos.

Para deshacer ese equívoco tan arraigado en nuestros investigadores la clave está en tomar de la vida común los razonamientos lógicos que usamos en situaciones equivalentes a estas y ver que son válidos al intentar determinar el tamaño idóneo de muestra en la investigación científica.

a) Chin Chu Li pregunta cuál es la cantidad de dinero adecuada

Hallábanse Aída y el coronel Gerineldo Márquez reunidos con sus amigos Prudencio Aguilar y Pilar Ternera criticando duramente la insensatez de los estadísticos, que responden con cuatro números totalmente diferentes cuando se les pide uno solo, cuando acertó a pasar por allí un bibliotecario (en el sentido borjiano de la palabra) chino que dijo llamarse Chin Chu Li.

Acababa de llegar a Occidente y proyectaba ir a pasar unos días a la feria de Sevilla y no conociendo nada acerca de ese país y su moneda, pregunta, a cada uno de los cuatro asistentes por separado, cuál es la cantidad «adecuada» de euros que debe llevar para ese viaje. He aquí las respuestas que cada uno le dio, tras hablar con él acerca de su proyecto.

Aída Buendía . . . 100 euros;	Gerineldo Márquez 700 euros
Pilar Ternera . . . 1.500 euros;	Prudencio Aguilar 4.000 euros

A la vista de esas divergencias, Chin les reprocha enérgicamente sus contradicciones y amenaza con hacerse el arakiri de inmediato si no se le da un número y solo uno: «el adecuado, el correcto, el verdadero...», clama Chin desconsoladamente.

Pero sus amigos le explican que no hay contradicción alguna entre esas cuatro diferentes cantidades, pues cada una de ellas es «correcta», según las comodidades que desee disfrutar. Concretamente cada uno de ellos calculó la cantidad de dinero de acuerdo con estos requerimientos:

Aída Buendía (100 E)	→ Dos días en pensión modesta, comiendo bocadillos.
Gerineldo Márquez (700 E)	→ Cuatro días en hotel modesto, restaurantes populares.
Pilar Ternera (1.500 E)	→ Seis días en hotel bueno, restaurantes confortables.

Prudencio Aguilar (4.000 E) → Diez días, hotel y restaurantes con lujo asiático.

Al serle explicados los motivos de aquellas diferencias, el doctor Chin ya no insiste más en que se le de un solo número, pues comprende que muy diferentes cantidades de dinero pueden ser «adecuadas». No cabe calcular la cantidad «correcta», sino la cantidad «adecuada a las comodidades que desee disfrutar», y ese nivel de comodidad *tiene que decidirlo él*, no sus asesores.

No obstante, hay cantidades claramente inadecuadas, por ser demasiado pequeñas o demasiado grandes. Por ejemplo, 10 euros serían totalmente insuficientes y un millón sería manifiestamente excesivo.

Organizan una gran fiesta para celebrar que están todos de acuerdo y tienen tan claras estas *sencillas ideas*. Y todos coinciden en que las entendería perfectamente un niño de 10 años y una persona sin ningún tipo de estudios. El visitante repite constantemente y les hace repetir a ellos con sospechosa reiteración aquello que todos tienen totalmente claro:

«Muy poco o muchísimo dinero sería claramente inadecuado, pero excluidos esos extremos, cualquier cantidad es válida porque permite asistir a la Feria, aunque la estancia será más prolongada y cómoda cuanto mayor sea la cantidad de dinero».

Tantas veces les hace repetir el visitante ese estribillo que deciden preguntarle la causa de su excesiva reiteración y él les responde que pronto se la dirá pero antes tiene que pedirles asesoramiento sobre otra cuestión que le preocupa.

b) Chin Chu Li pregunta cuál es la cantidad de tiempo adecuada

Y así les pregunta, nuevamente a cada uno por separado, cuánto tiempo debe quedarse a vivir en España para aprender español. He aquí las respuestas que le dieron tres de sus interlocutores, tras hablar con él acerca del tema:

- Aída Buendía: 5 años.
- Pilar Ternera: 2 años.
- Prudencio Aguilar: 4 meses.

Y nuevamente Chin les reprocha enérgicamente sus contradicciones y amenaza con apocalípticos horrores si no se le da un número y solo uno: «el adecuado, el correcto, el verdadero...», clama Chin amargamente.

Y sus amigos se ven obligados a explicarle una vez más que no hay contradicción alguna entre esas tres diferentes cantidades, pues cada una de ellas es «correcta», según los aspectos del idioma que desee aprender. Concretamente cada uno calculó el tiempo de acuerdo con estos requerimientos:

- Aída Buendía (5 años) → Dominio de la filología española y pronunciación cuasi perfecta.
- Pilar Ternera (2 años) → Hablar y escribir con soltura, pronunciación aceptable.
- Prudencio Aguilar (4 meses) → Nociones básicas para sobrevivir, pronunciación pintoresca.

Y también ahora al serle explicados los motivos de aquellas diferencias el doctor Chin ya no insiste en que se le dé un solo número, pues comprende que muy diferentes cantidades de tiempo pueden ser «adecuadas». No cabe calcular el tiempo «correcto», sino el tiempo «adecuado a la cantidad de conocimiento que desee adquirir», y ese nivel de conocimiento *tiene que decidirlo él*, no sus asesores.

No obstante, hay periodos de tiempo claramente inadecuados, por ser demasiado pequeños o demasiado grandes. Por ejemplo, 10 días serían totalmente insuficientes y 20 años serían manifiestamente excesivos.

Y cuando Chin comprende esto organizan otra gran fiesta para celebrar que están todos de acuerdo y tienen tan claras estas *sencillas ideas*. Y todos coinciden en que las entendería perfectamente un niño de 8 años y una persona analfabeta. Pero también ahora el visitante repite constantemente y les hace repetir a ellos con sospechosa reiteración aquello que todos tienen totalmente claro:

«Muy poco o muchísimo tiempo sería claramente inadecuado, pero excluidos esos extremos cualquier cantidad es válida porque permite aprender algo de español, aunque cuanto más tiempo dure el aprendizaje más se conocerá el idioma».

Nuestros protagonistas están cada vez más sorprendidos con el comportamiento del invitado. ¿Cómo es posible que en el segundo de los epi-

sodios no tuviera ya bien claro que no hay la «cantidad adecuada» que buscaba, sino que esta depende de lo que se quiera conseguir? ¿Por qué esa insistencia en repetir una y otra vez esas frases tan obvias que no parece necesario decirlas más de una vez? Al fin deciden preguntarle de nuevo la causa de su excesiva reiteración y esta vez no le dejan opción a posponer la respuesta. Y al fin el visitante confiesa la verdad.

c) Chin Chu Li desvela las causas de su presencia

Aunque de origen chino es ciudadano americano, autor de un libro (delicioso, ciertamente) sobre análisis de la varianza y ha venido, a requerimiento de los más cualificados estadísticos, para intentar deshacer de una vez por todas los nefastos malentendidos que hay entre los investigadores biológicos sobre el «tamaño idóneo de la muestra».

Les explica que la desesperación de Aída y de Gerineldo, porque les daban diferentes tamaños de muestra, son tan injustificadas como la suya cuando le indicaban distintas cantidades de dinero para gastar en la Feria o diferentes periodos de tiempo para aprender un idioma.

Porque excluidos los muy pequeños o muy grandes cualesquiera de los tamaños son válidos, ya que proporcionarán información útil, si bien, en general, cuanto mayor sea la muestra más información aportará (de la misma forma que cuanto más dinero emplee más comodidades y días tendrá en la Feria y cuanto más tiempo dedique a estudiarlo, mejor conocerá un idioma).

Pero esa relación —a más tamaño, más información— crece de modo continuo y no hay, en general, un tamaño que marque un cambio cualitativo.

Así, los 4 tamaños que le proponen a Aída (50, 374, 1.040 y 27.225 individuos en la muestra) son válidos, pero los mayores le darán más información. Concretamente:

- Si toma una muestra de $N = 50$ tiene probabilidad = 0,90 de que la proporción poblacional no se aleje de la encontrada en la muestra en más de 0,07, si la proporción poblacional es del orden de 0,10 o de 0,90.
- Si toma una muestra de $N = 374$ tiene probabilidad = 0,99 de que la proporción poblacional no se aleje de la encontrada en la muestra en más de 0,04, si la proporción poblacional es del orden de 0,10 o de 0,90.

- Si toma una muestra de $N = 1.040$ tiene probabilidad = 0,99 de que la proporción poblacional no se aleje de la encontrada en la muestra en más de 0,04, si la proporción poblacional es del orden del 0,50.
- Si toma una muestra de $N = 27.225$ tiene probabilidad = 0,999 de que la proporción poblacional no se aleje de la encontrada en la muestra en más de 0,01, si la proporción poblacional es del orden de 0,50.

Por tanto, cuanto mayor es la muestra, más probabilidad se tiene de que la proporción muestral esté más cerca de la poblacional. Ello varía gradualmente y no hay un valor de corte que marque una frontera.

También eran válidos cada uno de los tamaños que le dieron a Gerineldo Márquez: 11, 83, 265 y 1.357. Si hubiera convenido declarar que «B» es mejor que «A» cuando al hacer el test estadístico se encuentra $P < 0,05$, tendría, en efecto, una potencia de 0,85%.

- ... con muestras de $N = 11$ en cada grupo si A evita realmente a 80%.
- ... con muestras de $N = 83$ en cada grupo si A evita realmente a 50%.
- ... con muestras de $N = 265$ en cada grupo si A evita realmente a 40%.
- ... con muestras de $N = 1.357$ en cada grupo si A evita realmente a 35%.

Es decir, para que en 85 de cada 100 estudios aparezca $P < 0,05$ debe usar 11 ratas en cada grupo si realmente A evita el 80% de cánceres, pero si A evita el 35% de cánceres, necesita 1.357 ratas por grupo. Por tanto, cuanto menor es el efecto real, mayor tamaño de muestra se necesita para tener la misma probabilidad de detectar que hay ese efecto.

TAMAÑO DE MUESTRA ADECUADO PARA ESTIMAR UNA PROPORCIÓN POBLACIONAL

Cuando se le explica al investigador que, en contra de su creencia, no hay un tamaño que sea el «adecuado» para la investigación que proyecta, suele responder que en los libros de estadística aparecen fórmulas destinadas a calcular esos tamaños y él cree que aplicando esa fórmula a su proyecto actual aparecería el citado tamaño idóneo. En este apartado aplicamos la fórmula correspondiente al proyecto de Aída y veremos por qué pueden obtenerse muchas respuestas distintas, todas ellas válidas. La clave está en que el tamaño de muestra que proporciona esa fórmula depende de ciertas cantidades.

Queremos conocer el porcentaje poblacional de consumidores de hipnóticos (CH). Dado que no podemos estudiar toda la población analizaremos una muestra. ¿De qué tamaño? Es obvio que ningún tamaño muestral permite conocer exactamente el dato poblacional. A partir del porcentaje muestral (%_{MUES}) calcularemos un *intervalo de confianza (IC)* dentro del cual tenemos cierta confianza en que se encuentre el porcentaje poblacional (%_{POBL}).

En general, cuanto mayor sea la muestra más precisa será la estimación que hagamos, es decir el IC para el valor poblacional, IC (%_{POBL}), será más estrecho. Veamos estos ejemplos:

Con N = 20 y 2 CH: %_{MUES} = 10% → IC_{95%} (%_{POBL}) = 0,1% y 23%
Anchura IC: 22,9

Con N = 400 y 40 CH: %_{MUES} = 10% → IC_{95%} (%_{POBL}) = 7% y 13%
Anchura IC: 6

Con N = 1.600 y 160 CH: %_{MUES} = 10% → IC_{95%} (%_{POBL}) = 8,5% y 11,5%
Anchura IC: 3

Con N = 1600 y 160 CH: %_{MUES} = 10% → IC_{99%} (%_{POBL}) = 8% y 12%
Anchura IC: 4

Vemos que la precisión de la estimación, es decir la anchura del IC, depende del tamaño de muestra y de la confianza que queremos para el intervalo. No hay un tamaño idóneo, sino que cada tamaño proporciona cierta confianza en que el error de estimación (distancia entre la proporción poblacional y la muestral) no supere cierta cantidad.

El tamaño para estimar la proporción poblacional, N, se calcula con esta fórmula:

$$N = p'(1-p') Z_a^2 / d^2$$

Vemos que N depende de tres cantidades, cuyo significado explicamos a continuación:

p' → Indica el orden de magnitud de la proporción poblacional que intentamos conocer a través del estudio que estamos proyectando.

Observación: esa proporción no se conoce, por eso se va a hacer un estudio para averiguar cuánto vale. La cantidad p' a poner en la fórmula debe ser una que suponemos estará cercana a la proporción que intentamos averiguar. Si no hay ninguna información al respecto, ponemos $p' = 0,5$.

d → Queremos error de estimación menor de «d», es decir, que la proporción que aparezca en la muestra no se aleje de la verdadera proporción poblacional en más de «d»: $|\Pi - p| < d$. Esta cantidad debe decidirla el investigador y no hay criterios «matemáticos» ni «científicos» para hacerlo. Obviamente, según el valor que se ponga para «d» va a salir un tamaño u otro y el tamaño que sale con cada valor de «d» que se ponga es el adecuado, en el sentido que luego veremos, para ese error.

α → Confianza en que vaya a cumplirse que $|\Pi - p| < d$. Una vez que el investigador haya decidido el valor «d» que quiere usar, hay que decirle que no hay ningún tamaño que garantice totalmente el cumplimiento de esa condición, es decir, que sea $|\Pi - p| < d$. Una vez más, la confianza que el investigador tiene en que ocurra eso es mayor cuanto mayor es la muestra. El valor « Z_α » depende de cuán grande queremos que sea esa confianza. Concretamente vale 1,65 para confianza de 90%, 1,96 para confianza de 95%, 2,58 para confianza de 99% y 3,3 para confianza de 99,9%.

Apliquemos la fórmula en primer lugar para el caso en que Aída dijera que quería tener 99% en que la verdadera proporción poblacional de CH no se aleje más de 4 puntos ($d = 0,04$) de la que aparezca en la muestra. Y supongamos que ella considera que la proporción poblacional puede ser del orden del 10%. Es decir: $p' = 0,10$, $d = 0,04$ y $Z_\alpha = 2,58$.

$$N = (0,1 \times 0,9) 2,58^2 / 0,04^2 = 374$$

Los valores que Aída decide son razonables, como también otros muchos posibles y no hay razón objetiva que haga más válidas esas cantidades que otras parecidas. Y con otras especificaciones, es decir, otros valores de confianza, error máximo deseable y p' , el resultado es distinto. Aquí están las distintas especificaciones y tamaños:

Confianza	d	p'	N
90%	.07	.10	$1,65^2 (0,1 \times 0,9) / 0,07^2 = 50$
99%	.04	.10	$2,58^2 (0,1 \times 0,9) / 0,04^2 = 374$
99%	.04	.50	$2,58^2 (0,5 \times 0,5) / 0,04^2 = 1.040$
99,9%	.01	.50	$3,3^2 (0,5 \times 0,5) / 0,01^2 = 27.225$

Es decir, si tomo una muestra de $N = 50$ tengo probabilidad = 0,90 de que la proporción poblacional no se aleje de la encontrada en la muestra

en más de 0,07, si la proporción poblacional es del orden del 0,10. Pero si tomo una muestra de $N = 27,225$ tengo probabilidad = 0,999 de que la proporción, poblacional no se aleje de la encontrada en la muestra en más de 0,01, si la proporción poblacional es del orden del **0,50**.

EPÍLOGO

En todos los libros de Bioestadística encontrará las fórmulas para estimar tamaños de muestra. Pero lo relevante ahora no son esas fórmulas, sino el tomar conciencia de su limitada utilidad y poner fin al uso indiscriminado que frecuentemente se hace de ellas sin entender la información que nos dan.

En la inmensa mayoría de los casos el sentido común y los recursos disponibles sugieren tamaños de muestras válidos que el investigador debe usar sin complejos. Posteriormente, a la vista de los resultados obtenidos y de los recursos disponibles, decidirá si procede ampliar la muestra.

Las personas con poderes para aprobar o desestimar proyectos de investigación deben tener en cuenta esta realidad. Deben saber que no hay un «tamaño de muestra adecuado» ni un modo «científicamente correcto» de llegar a él. A igualdad de otras circunstancias, la cantidad de información obtenida en un estudio crece a medida que aumenta el tamaño de muestra. De modo orientativo, se puede hacer una estimación de las «prestaciones» que proporcionan algunos tamaños de muestra concretos, y en algunas ocasiones ello puede ayudar al investigador a elegir el tamaño de su estudio, teniendo en cuenta los recursos de que dispone.

La errónea idea de que cabe y debe calcularse con fórmulas estadísticas el tamaño de muestra adecuado está tan arraigada entre nuestros investigadores que es necesario insistir una vez más en cuál es la realidad.

«En la mayoría de las situaciones no ha lugar calcular el “tamaño adecuado”, porque NO hay UN solo tamaño que sea el adecuado. En general, ninguna muestra permitirá saber lo que ocurre exactamente en la población, y cuanto mayor sea la muestra más información proporcionará acerca de lo que realmente ocurre en la población, pero también consume más tiempo y recursos. El investigador debe decidir buscando un equilibrio entre los recursos que puede dedicar y la información que va a obtener».

Conclusiones

Podemos resumir lo visto en los últimos capítulos en los siguientes puntos:

1. El cálculo del valor P de los tests de significación (TS) es una cuestión matemática, pero entender lo que indica y usarlo razonadamente no requiere conocimientos matemáticos, es una cuestión de lógica básica al alcance de todas las personas, ya que *todas usan en la vida diaria ese mismo proceso lógico*.
2. Los TS nos ayudan en el proceso de ajustar nuestra opinión sobre una hipótesis (HP) a la vista de unos datos observados (DO).
3. Si los DO son incompatibles con la HP (porque no aparecen cuando es cierta), la rechazamos (puesto que han aparecido) y si son compatibles con la HP, aceptamos que puede ser cierta (pero no afirmamos que lo es, porque con esos DO también son compatibles otras HP).
4. Es manifiesta la asimetría de la situación:
 - a) DO incompatibles con la HP $\rightarrow$ aseguramos que es falsa.
 - b) DO compatibles con la HP $\rightarrow$ no aseguramos que es cierta, solo aceptamos que puede serlo.

Es decir, los DO nunca nos permitirán afirmar que una HP es cierta, pero puede que nos permitan afirmar que es falsa.

5. La pseudo-ciencia se caracteriza porque da por cierta una determinada HP entre las muchas con las que son compatibles los DO. La ciencia rigurosa nunca da por cierta una HP porque sean compatibles con ella los DO, si hay otras con las que también lo son. Más bien pone el acento en descartar las HP con las que son

incompatibles los DO, lo cual también implica avance en el conocimiento del tema estudiado.

6. El valor P del test nos dice con qué frecuencia relativa aparecen esos DO u otros más alejados de lo esperado bajo la hipótesis nula (H_0), cuando se toman muchas muestras de una población en la que es cierta la H_0 .
7. Cuanto menor es el valor P, más evidencia constituye contra la H_0 , pero no hay un valor frontera (ni el 5% ni el 1% ni ningún otro valor) que separe los valores de P que nos llevan a rechazar la H_0 de los que no permiten hacerlo. Esto no es un criterio específicamente estadístico, es un proceso lógico propio de la mente humana y que usamos continuamente en todos los órdenes de la vida. En la vida común manejamos constantemente magnitudes continuas en las que no hay un punto de corte que separe dos zonas conceptualmente distintas, aunque se pueda y se deba convenir un valor frontera con fines operativos prácticos.
8. La «regla del 5%», o cualquier otra cantidad convenida, es inadecuada para establecer conclusiones razonables en la investigación científica enfocada a la adquisición de conocimiento (no a la toma de decisiones).
9. Las expresiones «estadísticamente significativo» y «estadísticamente no significativo», son otro modo de decir « $P < 0,05$ » o « $P > 0,05$ » (u otro valor acordado) y por ello no aportan nada en el proceso de adquisición de conocimiento.
10. En la Toma de Decisiones sí es necesario decidir una cifra, que se suele llamar alfa, de modo que se ejecuta una u otra acción según que el valor de P sea mayor o menor que alfa.
11. La adquisición de conocimiento y la toma de decisiones son dos situaciones diferentes que requieren diferentes estrategias de inferencia estadística, tanto en la vida común como en la investigación científica. No constituyen estrategias contrapuestas, sino complementarias:
 - a) *Test de Significación* → para Formarse opinión, propio de la adquisición de conocimiento y

b) *Test de Hipótesis* → para *Elegir entre dos acciones*, propio de la *Toma de Decisiones*

12. Cada estudio considerado aisladamente no conduce, en muchos casos, a conclusiones definitivas y ni el autor ni sus lectores están obligados a decantarse a favor o en contra de una hipótesis cuando los resultados no lo permiten.
13. La comunidad científica no espera que el autor de un trabajo demuestre definitivamente que en la población hay o no hay cierto efecto. Lo correcto es dar el valor de P encontrado, de modo que el lector pueda valorar cuanta evidencia constituyen los datos contra la H_0 . Además, cuando es posible, deben darse los intervalos de confianza.
14. Que los experimentos con valor P no muy pequeño no permitan una conclusión segura, no implica que carezcan de valor y deban ser desechados, pues pueden constituir información muy útil al ser considerados *junto con otros trabajos* sobre el mismo tema. La comunidad científica rara vez se deja convencer por los resultados de un solo estudio. Por el contrario, tiene en cuenta *los resultados de estudios similares* y admite que cierto efecto es una realidad general cuando encuentra que ese efecto aparece reiteradamente en varias investigaciones.
15. Uno de los errores más frecuentes es creer que el valor P del test es la probabilidad de que sea cierta la hipótesis nula o la de trabajo. En la mayoría de los estudios realizados no hay modo de cuantificar esas probabilidades, por lo que la inferencia clásica se centra en evaluar la probabilidad de que aparezcan cierto tipo de muestras si es cierta la H_0 , es decir, el valor P del test.
16. Así como la evidencia contra la H_0 aumenta progresivamente al disminuir el valor P del test, sin valores que separen drásticamente dos zonas, también aumenta progresivamente la información que aporta la muestra estudiada. No hay un tamaño que sea el adecuado. Cuanto mayor sea la muestra más información proporcionará acerca de lo que realmente ocurre en la población.

Encuestas de autoevaluación previas

(Compruebe su nivel de conocimientos)

Muchos de los investigadores al publicar sus trabajos se atienen a ciertas «recetas» que suelen ser básicamente inadecuadas, pues la mayoría de las veces no constituyen una ayuda real para elaborar conclusiones razonables, sino más bien un artefacto que favorece el error.

Centenares de seminarios con los profesionales de Ciencias de la Salud ponen de manifiesto que la mayoría de los que decían en un cuestionario previo tener ideas claras sobre este tema, cometían errores importantes al ponerlas en práctica.

Para ayudarle a tomar conciencia de su nivel de conocimiento, aquí tiene tres «encuestas previas» en las que se le invita a responder si cada una de las afirmaciones es verdadera o falsa.

Marque cada una de las afirmaciones que allí aparecen con una «V» si cree que es verdadera (debe ser cierta toda la frase en su conjunto, no solo una parte de ella tomada aisladamente) y con una «F» si cree que es falsa, absurda, ininteligible o inadecuada. Si no lo sabe no la califique, pues en la puntuación se penalizan más las respuestas equivocadas que las abstenciones.

En el Apéndice 3 se le dan las respuestas correctas para que usted pueda autoevaluarse. Le proponemos que dedique una sesión tranquila a responderlas y luego verifique su nivel de aciertos y errores. Es muy probable que descubra que ha cometido más errores de lo que usted suponía. Si es así esperamos que la lectura atenta de este libro le ayude decisivamente a aclarar algunas ideas fundamentales.

Todas las afirmaciones propuestas en estas encuestas han sido contrastadas varias veces, de modo que las erratas formales o imprecisiones conceptuales son mínimos. Si hay muchas afirmaciones que a usted le parecen un “galimatías” o “complicados juegos de palabras” es que necesita revisar sus conceptos básicos sobre este tema.

En las encuestas que a continuación le proponemos tenga en cuenta que si para cierta afirmación son válidas las expresiones «*Es casi seguro*» y «*Es posible*», el «*Es posible*» debe considerarlo como inadecuado. En la vida común diferenciamos claramente estas dos opciones y es imprescindible hacer esa misma distinción al elaborar las conclusiones de los trabajos científicos. Por ejemplo: la afirmación «Si conducimos en ciudad a 200 km/h y pasando los semáforos en rojo *es posible* que tengamos un accidente», la daríamos como inadecuada, porque lo adecuado sería: «*Es casi seguro* que tendremos un accidente».

Por el contrario, si decimos «*Es casi seguro* que en el decenio 2004-2013 habrá un Premio Nobel español», lo consideramos inadecuado, y lo adecuado sería «*Es posible* que en el decenio 2004-2013 haya...».

MODO DE PUNTUAR LAS ENCUESTAS

Por cada respuesta correcta sume un punto, reste uno por cada una equivocada, y no ponga ni quite puntos por las frases en las que usted no se ha pronunciado. Con este criterio usted sumaría tantos puntos como afirmaciones hay en la encuesta si contesta todo correcto, tendrá «menos esa cantidad de puntos» si contesta equivocadamente todas las preguntas y tendrá «cero» si tiene tantos aciertos como fallos, que es lo que esperaríamos en una persona que desconociera totalmente el tema y respondiera al azar.

Una vez tenga la puntuación alcanzada por usted póngala en «escala de 0 a 10», para lo cual tiene que dividirla por el número de afirmaciones que tiene la encuesta y multiplicar por 10. Este número, que irá desde -10 en los que contesten todo equivocado, hasta 10 en los que contesten todo correcto, es lo que llamamos «nivel» obtenido en cada encuesta.

Un nivel de 9 o más, indica que se conoce bien esta parte de la materia (o se ha tenido suerte extraordinaria al responder). De 7 a 8,9 indicaría que hay alguna laguna seria y es preciso revisar los conceptos en los que se cometieron los fallos. Menos de 7 sugiere necesidad de releer todo el capítulo muy atentamente, sin prisas y sin pausas.

ENCUESTA DE AUTOEVALUACIÓN PREVIA-1

Para estudiar el posible efecto anticancerígeno (AC) de 2 productos, «A» y «B», trabajaremos con una cepa de ratas genéticamente modificada, en la que el 90% de ellas desarrollan cáncer espontáneamente el segundo año de su vida.

Se prueba cada producto en 20 ratas. He aquí los resultados y el valor P del test de significación para cada fármaco así como los intervalos de confianza (IC) para el % de cánceres que se obtendría al dar el fármaco a toda la población de este tipo de ratas.

«A» → Hacen cáncer 8 ratas → 40%, $P = 0,000003$, $IC_{95\%} = 19\%$ y 64%

«B» → Hacen cáncer 18 ratas → 90%, $P = 0,608$, $IC_{95\%} = 69\%$ y 99%

1. Es casi seguro que «A» es AC.
2. Es posible que «A» sea AC.
3. Es casi seguro que con «A» el % poblacional de cánceres es 40%.
4. Es casi seguro que «A» es inútil.
5. Los datos son compatibles con que «A» sea inútil.
6. Lo razonable es concluir que «A» es inútil.
7. Si «A» fuera inútil, en un millón de estudios como este solo 3 darían 8 o menos cánceres.
8. Es casi seguro que «B» es AC.
9. Es posible que «B» sea AC.
10. Es casi seguro que administrando «B» el % poblacional de cánceres es 90%.
11. Es casi seguro que «B» es inútil.
12. Los datos son compatibles con que «B» sea inútil.
13. Lo razonable es concluir que «B» es inútil.
14. Si «B» fuera inútil, en mil estudios como este 608 darían 18 o menos cánceres.

ENCUESTA DE AUTOEVALUACIÓN PREVIA-2

Se sospecha que el producto «A» puede ser teratógeno (produce malformaciones durante el desarrollo fetal) en ratas.

- *Entre 400 ratas nacidas de madres que habían recibido «A» hubo 48 con malformación: 12%.*
- *Entre 300 ratas nacidas de madres que NO habían recibido «A» hubo 6 con malformación: 2%.*

Por tanto, en las muestras el riesgo relativo fue: $RR = 12/2 = 6$. Es decir, el % de malformaciones es 6 veces mayor en la muestra con «A» que en la muestra sin «A». Para ver si este efecto es una realidad más allá del caso particular de las muestras estudiadas hacemos el test estadístico y se encuentra $P_{UNILATERAL} = 0,000\ 000\ 4$.

1. La hipótesis nula establece que «A» es teratógeno, es decir, $RR_{Poblacional} > 1$.
2. La hipótesis nula plantea que «A» no es teratógeno, es decir, $RR_{Poblacional} = 1$.
3. El efecto encontrado en la muestra ($RR_{Muestral} = 6$) también existe en la población y es de la misma magnitud ($RR_{Poblacional} = 6$).
4. Si «A» es teratógeno es muy difícil encontrar en la muestra estudiada un RR en torno a 6.
5. Si «A» no es teratógeno es muy difícil encontrar en la muestra estudiada un RR en torno a 6.
6. Lo razonable, a la vista del resultado, es pensar que «A» es teratógeno.
7. El resultado es claramente compatible con que «A» no sea teratógeno.
8. Siendo muy difícil que este tipo de resultado ($RR_{muestral} = 6$) aparezca por azar, no constituye evidencia clara a favor de que «A» es teratógeno.
9. Siendo muy difícil que este tipo de resultado ($RR_{muestral} = 6$) aparezca por azar, constituye evidencia clara a favor de que «A» es teratógeno.
10. El pequeño valor P obtenido indica que muy probablemente en la población el RR es 6.

11. Si «A» es teratígeno, en 10 millones de estudios como este solo 4 darían un valor de RR *muestral* como 6 o mayor.
12. Si «A» no es teratígeno, en 10 millones de estudios como este solo 4 darían un valor de RR *muestral* como 6 o mayor.

ENCUESTA DE AUTOEVALUACIÓN PREVIA-3

Se sospecha que el producto «A» puede bajar la tensión arterial (TA). En una muestra de 10 perros se mide la TA antes y después de darles «A» en las dosis establecidas. En esa muestra se encuentra un descenso medio de 20 mm Hg, con error estándar de 15, y al hacer el test estadístico se encuentra $P_{\text{UNILATERAL}} = 0,12$.

1. La *hipótesis nula* establece que «A» no modifica la TA en la población, es decir que la media de la TA en la población antes y después de recibir «A» es la misma.
2. La *hipótesis nula* establece que «A» modifica la TA en la población, bajándola en 20 mm Hg.
3. El resultado es claramente compatible con que «A» no sea hipotensor.
4. El resultado es difícilmente compatible con que «A» no sea hipotensor.
5. Puesto que P es grande, se demuestra que «A» no es hipotensor.
6. Hay una probabilidad de un 12% de que «A» sea hipotensor.
7. Hay una probabilidad de un 12% de que «A» no sea hipotensor.
8. Siendo *fácil* que un resultado de este tipo (descenso medio muestral de 20 mm Hg) aparezca por azar, no constituye evidencia clara a favor de que «A» es hipotensor.
9. Es muy probable que realmente el producto «A» baje la TA en 20 unidades por término medio, es decir, que si se diera a toda la población la media de la TA bajara en 20 unidades.
10. Si «A» *no* es hipotensor, en 100 estudios como este 12 darían un descenso medio muestral como 20 o mayor.
11. Si «A» es hipotensor, en 100 estudios como este 12 darían un descenso medio muestral como 20 o menor.

Encuestas de autoevaluación específicas

A continuación se incluyen autoevaluaciones específicas para varios capítulos. Se recomienda realizar cada una al acabar el capítulo correspondiente y no pasar al siguiente si la puntuación (obtenida con el criterio especificado en el Apéndice 1) es inferior a 9.

Marque cada una de las siguientes afirmaciones con una «V» si cree que es verdadera (debe ser cierta toda la frase en su conjunto, no solo una parte de ella tomada aisladamente) y con una «F» si cree que es falsa, absurda, ininteligible o inadecuada. Si no lo sabe no la califique, pues en la puntuación se penalizan más las respuestas equivocadas que las abstenciones.

ENCUESTA DE AUTOEVALUACIÓN DEL CAPÍTULO 4

Como gobernador de la Ínsula de Babaria, Sancho Panza es responsable de su Casino, en cada una de cuyas 4 mesas se juegan inmensas fortunas mediante el lanzamiento de una moneda. El juego se establece asumiendo que cada una de esas 4 monedas es perfectamente equilibrada, de modo que al lanzarla más y más veces la proporción de lanzamientos en que sale «cara» se aproxima más y más al 50%.

Algunos resultados recientes hacen sospechar que algunas de las monedas podrían haber sido trucadas por el correspondiente crupier para que salgan más del 50% de caras. Ante la denuncia, Sancho expone

que lo razonable es partir de la hipótesis de inocencia de cada uno de los crupieres, es decir, que no han trucado las monedas. Esa hipótesis solo se abandonará si aparecen resultados incompatibles, o muy difícilmente compatibles, con que la moneda sea equilibrada.

La «investigación» consistirá en lanzar cada una de las cuatro monedas cierto número de veces y ver si el número de caras que aparecen es razonablemente próximo al 50%, o, por el contrario, extremadamente alejado de esa cantidad.

Estos son los resultados obtenidos al investigar cada moneda:

Moneda	Número de lanzamientos	Número de caras	Porcentaje de caras
A	200	200	100%
B	3	3	100%
C	50	27	54%
D	30	15	50%

Hagamos un TS (Test de Significación) para elaborar conclusiones razonables respecto a la moneda «A»

1. La hipótesis nula, H_0 , planteada dice que la moneda «A» es equilibrada.
2. A la vista del resultado concluimos que la moneda «A» *no es equilibrada*, es decir, si se lanzara muchas veces a la larga saldría cara en más de la mitad de los lanzamientos: rechazo la hipótesis de inocencia.
3. A la vista del resultado concluimos que la moneda «A» *puede que sea equilibrada*, es decir, acepto la hipótesis de inocencia como posible.
4. A la vista del resultado concluimos que la moneda «A» *es equilibrada*, es decir, afirmo que la hipótesis de inocencia es cierta.
5. Si «A» fuera equilibrada sería muy difícil que al tirarla 200 veces salieran todo caras, por eso creemos que *no es equilibrada*.
6. Si «A» fuera equilibrada sería muy fácil que al tirarla 200 veces salieran todo caras, por eso creemos que el hecho de que hayan salido 200 caras seguidas es *compatible con que sea equilibrada*.

Hagamos un TS para elaborar conclusiones razonables respecto a la moneda «B»

7. La hipótesis nula, H_0 , planteada dice que la moneda «B» *no es equilibrada*.
8. A la vista del resultado concluimos que la moneda «B» *no es equilibrada*, es decir, rechazo la hipótesis de inocencia.
9. A la vista del resultado concluimos que la moneda «B» *puede que sea equilibrada*, es decir, acepto la hipótesis de inocencia como posible.
10. A la vista del resultado concluimos que la moneda «B» *es equilibrada*, es decir, *afirmo* que la hipótesis de inocencia es cierta.
11. Si «B» fuera equilibrada sería muy difícil que al tirarla 3 veces salieran todo caras, por eso creemos que *no es equilibrada*.
12. Si «B» fuera equilibrada sería fácil que al tirarla 3 veces salieran todo caras, por eso creemos que el hecho de que hayan salido 3 caras seguidas *es compatible con que sea equilibrada*.

Hagamos un TS para elaborar conclusiones razonables respecto a la moneda «C»

13. La hipótesis nula, H_0 , planteada dice que la moneda «C» *está trucada*.
14. A la vista del resultado concluimos que la moneda «C» *no es equilibrada*, es decir, rechazo la hipótesis de inocencia.
15. A la vista del resultado concluimos que la moneda «C» *puede que sea equilibrada*, es decir, acepto la hipótesis de inocencia como posible.
16. A la vista del resultado concluimos que la moneda «C» *es equilibrada*, es decir, *afirmo* que la hipótesis de inocencia es cierta.
17. Si «C» fuera equilibrada sería muy difícil que al tirarla 50 veces salieran 27 caras, por eso creemos que *no es equilibrada*.
18. Si «C» fuera equilibrada sería muy fácil que al tirarla 50 veces salieran 27 caras, por eso creemos que el hecho de que hayan salido 27 caras en los 50 lanzamientos *es compatible con que sea equilibrada*.

Hagamos un TS para elaborar conclusiones razonables respecto a la moneda «D»

19. La hipótesis nula, H_0 , planteada dice que la moneda «D» *es equilibrada*.
20. A la vista del resultado concluimos que la moneda «D» *no es equilibrada*, es decir, rechazo la hipótesis de inocencia.
21. A la vista del resultado concluimos que la moneda «D» *puede que sea equilibrada*, es decir, acepto la hipótesis de inocencia como posible.
22. A la vista del resultado concluimos que la moneda «D» *es equilibrada*, es decir, *afirmo* que la hipótesis de inocencia es cierta.
23. Si «D» fuera equilibrada sería muy difícil que al tirarla 30 veces salieran 15 caras, por eso creemos que *no es equilibrada*.
24. Si «D» fuera equilibrada sería muy fácil que al tirarla 30 veces salieran 15 caras, por eso creemos que el hecho de que hayan salido 15 caras en los 30 lanzamientos nos asegura que *es equilibrada*.

ENCUESTA DE AUTOEVALUACIÓN DEL CAPÍTULO 5

Sospechando que el ejercicio moderado (EM) ayuda a bajar el nivel de colesterol total en plasma (CT), hacemos un estudio con 25 adultos que tienen una media de CT de 300. Tras seis meses caminando rápido media hora al día les volvemos a medir el CT.

1. La hipótesis de trabajo es que el EM no baja el CT.
2. La hipótesis nula planteada es que el EM no baja el CT.
3. La afirmación: «El EM baja el CT» quiere decir que si todos los pacientes con CT elevado hicieran EM, la media de CT de esa población sería más baja que antes del ejercicio.
4. Observando cómo varía la media de CT en una muestra tras 6 meses de hacer EM, puedo saber cómo habría variado la media de la población a la que pertenece esa muestra.
5. Conocer cómo varía la media de CT en una muestra tras 6 meses de hacer EM no nos permite saber cuánto habría variado exactamente la media de la población a la que pertenece esa muestra.
6. Conocer cómo varía la media de CT en una muestra tras 6 meses de hacer EM nos permitirá calcular un intervalo de confianza para la variación de la media poblacional correspondiente.
7. Conocer cómo varía la media de CT en una muestra tras 6 meses de hacer EM, puede que nos permita saber que la media poblacional correspondiente se modifica con el EM, aunque no podamos saber cuánto exactamente.
8. Si tras el EM la media muestral de CT baja 0,3 unidades, nos inclinamos a pensar que esa pequeña variación bien pudo ser por azar y no es evidencia a favor de que el EM realmente baja el CT.
9. Si tras el EM la media muestral de CT baja 120 unidades, concluimos que esa notable variación es evidencia fuerte a favor de que el EM realmente baja el CT, pues si el EM no tuviera efecto sería muy difícil que por puro azar hubiera una bajada tan acusada.
10. Si tras el EM la media muestral de CT baja 15 unidades, no está claro si puede ser una variación casual o más bien hay que atribuirle al efecto del EM.

Un sinólogo sospecha que últimamente nacen en China más varones que mujeres y para comprobarlo estudia una muestra de 200 partidas de nacimiento tomadas al azar entre los tres millones de nacimientos habidos el último año.

11. La hipótesis nula inicialmente planteada es que en la población de todos los recién nacidos chinos (RN) son varones la mitad.
12. La hipótesis nula inicialmente planteada es que en la población de todos los recién nacidos chinos (RN) son varones más de la mitad.
13. Se tenderá a rechazar la H_0 que dice que son varones el 50% de los RN si en la muestra son varones muchos más de la mitad.
14. Se tenderá a rechazar la H_0 que dice que son varones el 50% de los RN si el porcentaje de varones en la muestra es próximo al 50%.
15. Si en la muestra son varones el 87%, rechazaremos la H_0 y concluiremos que en la población son varones más del 50%.
16. Si en la muestra son varones el 87%, rechazaremos la H_0 y concluiremos que en la población son varones el 87%.
17. Si en la muestra son varones el 87%, pensaremos que la H_0 puede ser cierta, porque ese dato muestral es claramente compatible con que en la población sean varones el 50%.
18. Si en la muestra son varones el 51,5%, pensaremos que la H_0 puede ser cierta, porque ese dato muestral es compatible con que en la población sean varones el 50%.
19. Si en la muestra son varones el 51,5%, rechazaremos la H_0 y concluiremos que en la población son varones el 51,5%.
20. Si en la muestra son varones el 51,5%, rechazaremos la H_0 y concluiremos que en la población son varones más del 50%.

ENCUESTA DE AUTOEVALUACIÓN DEL CAPÍTULO 7

En un tiempo ya pretérito la lotería de Babilonia se ejecutó usando tres bombos, A, B y C, cada uno de los cuales contenía millones de bolas homogéneas, 10% de ellas blancas. J. L. B. sospechó que quizá en alguno de ellos se había aumentado la proporción de bolas blancas, de modo que ya no sería 10%, sino más.

Para investigarlo se toma de cada bombo una muestra de 50 bolas al azar. Si el bombo es legal se espera que aparezcan en torno a 5 bolas blancas (5 es el 10% de 50). Si entre las 50 bolas de la muestra hay 6 o 7 blancas no se considera indicio claro de que se ha manipulado la población aumentando el número de blancas, pues esa cantidad puede aparecer fácilmente aunque la población contenga exactamente el 10%. Pero si encontramos que de las 50 bolas sacadas en la muestra son blancas, por ejemplo, 48, se considera este resultado claramente indicativo de que en la población hay más de 10% blancas.

Esta tabla recoge para cada uno de los bombos el número y porcentaje de bolas blancas, el valor P del test (para la H_0 : en el bombo hay 10% de bolas blancas) y el IC para la proporción poblacional, es decir, para la proporción de blancas en el bombo.

Bombo	Tamaño de la muestra	Número de bolas blancas	Porcentaje de blancas en la muestra	IC al 95% para el % real de blancas	Valor P unilateral del test
A	50	6	12%	5%-24%	0,38
B	50	10	20%	10%-34%	0,024
C	50	15	30%	18%-45%	0,00007

Test para el BOMBO «A»

1. Si de un bombo en el que hay realmente 10% de blancas extraemos muchas muestras al azar de $N = 50$ bolas, en 38 de cada 100 muestras habrá 10 o más blancas.
2. El resultado constituye fuerte evidencia a favor de que en el bombo *no* hay 10% blancas, sino bastante más porcentaje: rechazo la hipótesis H_0 .
3. Concluimos que en «A» *puede* que haya 10% blancas, es decir, aceptamos la H_0 como posible.

4. Concluimos que en «A» *hay* 10% de bolas blancas, es decir, afirmamos que la H_0 es cierta.
5. Si en «A» hay realmente 10% de bolas blancas, de cada 100 muestras de $N = 50$ que se tomen, en 38 habrá 6 *o más* bolas blancas.
6. Si en «A» hay realmente 10% de bolas blancas, de cada 100 muestras de $N = 50$ que se tomen, en 38 habrá 6 *o menos* bolas blancas.
7. Si en «A» hay realmente 12% de bolas blancas, de cada 100 muestras de $N = 50$ que se tomen, en 38 habrá 6 o más bolas blancas.
8. Si en «A» hay realmente *más de 10%* de bolas blancas, de cada 100 muestras de $N = 50$ que se tomen, en 38 habrá 6 o más bolas blancas.

Test para el BOMBO «B»

9. Si de un bombo en el que haya realmente 10% blancas extraemos muchas muestras al azar de $N = 50$, en 24 de cada 1.000 muestras habrá 10 o más bolas blancas.
10. La H_0 típica que se plantea es que en el bombo «B» hay realmente 20% de bolas blancas.
11. El resultado constituye fuerte evidencia a favor de que en el bombo *hay* 20% de bolas blancas.
12. Si en «B» hay realmente *más de 20%* de bolas blancas, de cada 1.000 muestras de $N = 50$ que se tomen, en 24 habrá 10 o más blancas.
13. Si en «B» hay realmente 10% de bolas blancas, de cada 1.000 muestras de $N = 50$ que se tomen, en 24 habrá 10 blancas.
14. Si en «B» hay realmente 10% de bolas blancas, de cada 1.000 muestras de $N = 50$ que se tomen, en 24 habrá 10 *o menos* blancas.
15. Si en «B» hay realmente 10% de bolas blancas, de cada 1.000 muestras de $N = 50$ que se tomen, en 24 habrá 5 o más blancas.

Test para el BOMBO «C»

16. Si de un bombo en el que haya realmente 10% de bolas blancas extraemos muchas muestras al azar de $N = 50$, en 7 de cada cien mil muestras habrá 15 o más bolas blancas.

17. La H_0 típica que se plantea es que en el bombo «C» hay realmente 30% de bolas blancas.
18. El resultado constituye fuerte evidencia a favor de que en el bombo *hay 30%* de bolas blancas.
19. El resultado constituye fuerte evidencia a favor de que en el bombo *hay más de 30%* de bolas blancas.
20. El resultado constituye fuerte evidencia a favor de que en el bombo *hay menos de 30%* de bolas blancas.
21. El resultado constituye fuerte evidencia a favor de que en el bombo «C» *hay 10%* de bolas blancas.
22. El resultado constituye fuerte evidencia a favor de que en el bombo «C» *hay más de 10%* de bolas blancas.
23. El resultado constituye fuerte evidencia a favor de que en el bombo «C» *hay menos de 10%* de bolas blancas.
24. El valor P de este test nos dice que si en «C» hay realmente 30% de bolas blancas, de cada cien mil muestras de $N = 50$ que se tomen, en 7 muestras habrá 15 o más bolas blancas.
25. El valor P de este test nos dice que si en «C» hay realmente 10% de bolas blancas, de cada cien mil muestras de $N = 50$ que se tomen, en 7 muestras habrá 15 o más bolas blancas.
26. El valor P de este test nos dice que si en «C» hay realmente 10% de bolas blancas, de cada cien mil muestras de $N = 50$ que se tomen, en 7 muestras habrá 15 o menos bolas blancas.

ENCUESTA DE AUTOEVALUACIÓN DEL CAPÍTULO 9

NOTA: el valor P de los tests mencionados en esta encuesta es unilateral si no se especifica lo contrario.

En una publicación sobre la efectividad del fármaco «A» aparece la siguiente frase: «En las muestras 'A' produjo un 20% más de curaciones que el estándar, con una significación estadística de $P=1\%$ ». Este es el resultado de un estudio en el que a una muestra de individuos se les trató con el fármaco estándar y otra con «A». La interpretación de esa frase es:

1. En la muestra de individuos tratados con «A» se produjo un 20% más de curaciones que en la muestra de tratados con el estándar.
2. Al hacer el test de significación correspondiente, la hipótesis nula planteada es: «A» cura 20% más que el estándar».
3. Al hacer el test de significación correspondiente, la hipótesis nula planteada es: «El fármaco «A» cura igual que el estándar».
4. En el 1% de los individuos tratados «A» resultó más efectivo.
5. El valor de $P = 1\%$ nos permite concluir que *realmente* «A» produce 20% más de curaciones que el estándar.

En una publicación sobre la efectividad del fármaco «B» aparece la siguiente frase: «En las muestras 'B' produjo un 30% más de curaciones que el estándar, pero el valor P del test fue $P = 27\%$, de modo que este resultado no es estadísticamente significativo».

6. Al hacer el test de significación correspondiente, la hipótesis nula planteada es: «B» es beneficioso, es decir, con él se curan más que con el estándar».
7. El estudio muestra que «B» no es útil, es decir, que con él no aumenta el número de curaciones.
8. En el 27% de los individuos tratados «B» resultó efectivo.
9. El valor de $P = 27\%$ nos permite concluir que *realmente* «B» no produce 30% más de curaciones que el estándar.
10. A la vista de los datos podemos asegurar que el porcentaje poblacional de curaciones con «B» es 30%.
11. Un resultado muestral como el obtenido o aún más extremo se produciría en el 27% de las muestras si «B» no fuera efectivo.

Dieta	N	Media	Error estándar
Vegetariana	20	180	50
Estándar	50	193	12

El TS da valor $P = 0,15$

Estos son los resultados de medir la concentración de colesterol total, CCT, en plasma en una muestra de $N = 20$ personas que siguen dieta vegetariana y en otra muestra de $N = 30$ personas con dieta estándar.

12. La hipótesis nula típica planteada dice que la dieta vegetariana no modifica la CCT.
13. Los datos sugieren que el 15% de las veces la dieta vegetariana baja la CCT.
14. A la vista del valor P encontrado, concluimos que la diferencia de CCT obtenida en la muestra ($193-180 = 13$) sugiere claramente que la dieta vegetariana disminuye la CCT.
15. En el 15% de las personas estudiadas la dieta vegetariana disminuyen la CCT.
16. Es posible que la bajada observada en la muestra ($193-180 = 13$) sea producto del azar y en realidad la dieta vegetariana no disminuya la CCT.
17. Una bajada de 13 unidades en la CCT habla claramente a favor de la efectividad de la dieta vegetariana para disminuir la CCT.

ENCUESTA DE AUTOEVALUACIÓN DEL CAPÍTULO 10

NOTA: algunas de las afirmaciones que se le proponen a continuación son realmente absurdas y a toda luces falsas. El lector puede preguntarse qué sentido tiene invitarle a confirmar obviedades elementales. La razón es que cuando elaboran las conclusiones de trabajos científicos, muchos profesionales hacen afirmaciones tan insostenibles como algunas de las aquí propuestas. Ver lo injustificado de esas afirmaciones en ejemplos de la vida común o de la actividad médica quizás le ayude a no hacerlas en la actividad investigadora.

La idea relevante de este capítulo es que ciertos datos constituyen una clara evidencia contra una hipótesis, mientras que otros no constituyen evidencia a su favor, sino que, simplemente, son compatibles con ella. Recuerde que en el diagnóstico médico se utiliza un proceso lógico en todo paralelo a los tests de significación.

Nos dicen que a la entrada del hospital está a punto de llegar en un taxi el nuevo director-gerente, de cuyo aspecto exterior no tenemos ninguna noticia. Vemos aproximarse un taxi y consideramos la hipótesis que dice que el viajero es el nuevo director.

1. Si el viajero del taxi es un joven de unos 16 años, rechazamos la hipótesis, es decir, pensamos que no es el nuevo director.
2. Si el viajero del taxi es un hombre de unos 45 años, afirmamos que la hipótesis es cierta, es decir, afirmamos que es el nuevo director.
3. Si el viajero del taxi es un hombre de unos 45 años, pensamos que la hipótesis puede ser cierta, es decir, pensamos que puede ser el nuevo director.
4. Si el viajero del taxi es un hombre de unos 30 años, nos inclinamos a pensar que no es el nuevo director, pero nos queda ciertas dudas al respecto.
5. Si el viajero del taxi es un anciano de unos 90 años, rechazamos la hipótesis con seguridad, es decir, pensamos que no es el nuevo director.
6. Si el viajero del taxi es un hombre de unos 70 años, nos inclinamos a pensar que no es el nuevo director, pero nos queda ciertas dudas al respecto.

7. Si llegan casi simultáneamente dos taxis y de uno baja un joven de unos 20 años y del otro baja un hombre de unos 50, pensamos que el joven no es el director y que el otro puede que lo sea.
8. Si llegan casi simultáneamente dos taxis y de uno baja un joven de unos 20 años y del otro baja un hombre de unos 50, pensamos que el joven no es el director y que el otro sí que lo es.
9. Si llegan casi simultáneamente dos taxis y de uno baja un hombre de unos 40 años y del otro baja un hombre de unos 50, pensamos que cualquiera de ellos puede ser el director.
10. Si llegan casi simultáneamente dos taxis y de uno baja un hombre de unos 40 años y del otro baja un hombre de unos 50, afirmamos que cada uno de ellos es el director.
11. Si llegan casi simultáneamente tres taxis y de uno baja un hombre de unos 40 años, del otro baja un hombre de unos 50 y del tercero uno de unos 60 años, afirmamos que cada uno de ellos es el director.
12. Si llegan casi simultáneamente tres taxis y de uno baja un hombre de unos 40 años, del otro baja un hombre de unos 50 y del tercero uno de unos 60 años, decimos que cada uno de ellos puede ser el director.

Un terapeuta de drogadictos se plantea la hipótesis que dice que su paciente Juan ha logrado desengancharse, ha superado la dependencia.

13. Si en la fiesta en que ambos coinciden como invitados ve a Juan esnifar compulsiva y reiteradamente, rechaza su hipótesis y concluye que el paciente sigue atrapado.
14. Si en la fiesta en que ambos coinciden como invitados ve que Juan rechaza sistemáticamente todas las invitaciones a esnifar, concluye que su hipótesis es cierta, es decir, que Juan ha superado la dependencia.
15. Si en la fiesta en que ambos coinciden como invitados ve que Juan rechaza sistemáticamente todas las invitaciones a esnifar, concluye que su hipótesis puede ser cierta, es decir, que lo observado es compatible con la hipótesis, pero no afirma que sea cierta.

ENCUESTA DE AUTOEVALUACIÓN DEL CAPÍTULO 11

Para estudiar el posible efecto anticancerígeno (AC) de 2 productos, «A» y «B», trabajaremos con una cepa de ratas genéticamente modificada, en la que el 90% de ellas desarrollan cáncer espontáneamente el segundo año de su vida.

Se prueba cada producto en 20 ratas. He aquí los resultados y el valor P del test de significación para cada uno de los fármacos, así como los intervalos de confianza (IC) para el % de cánceres que se obtendría al dar el fármaco a toda la población de este tipo de ratas:

«A» → Hacen cáncer 15 ratas → 75%, $P_{\text{UNIL}} = 0,043$, $IC_{95\%} = 51\%$ y 91%
 «B» → Hacen cáncer 16 ratas → 80%, $P_{\text{UNIL}} = 0,133$, $IC_{95\%} = 56\%$ y 94%

16. Es casi seguro que «A» es AC.
17. Es casi seguro que administrando «A» el % poblacional de cánceres es 75%.
18. Es casi seguro que «A» es inútil.
19. Los datos son compatibles con que «A» sea inútil.
20. Lo razonable es concluir que «A» es útil.
21. Si «A» fuera inútil, en mil estudios como este 43 darían 15 o menos cánceres.
22. Es casi seguro que «B» es AC.
23. Es casi seguro que administrando «B» el % poblacional de cánceres es 80%.
24. Es casi seguro que «B» es inútil.
25. Los datos son compatibles con que «B» sea inútil.
26. Lo razonable es concluir que «B» es inútil.
27. Si «B» fuera inútil, en mil estudios como este 133 darían 16 o menos cánceres.
28. Las conclusiones habrán de ser claramente distintas para «A» y «B» puesto que los respectivos valores de P se encuentran a distinto lado de la frontera del 5%.
29. En resumen, concluiríamos que «A» es AC y «B» no lo es.
30. En resumen, concluimos que «A» es AC y «B» puede serlo o no serlo.
31. En resumen, concluimos que «A» puede ser o no ser AC y «B» no lo es.

-
32. En resumen, concluimos que ninguno de los productos es AC.
 33. En resumen, concluimos que cada uno de los dos productos pueden ser AC o no serlo.
 34. Los datos no permiten pronunciarse respecto a ninguno de los dos fármacos.

ENCUESTA DE AUTOEVALUACIÓN DEL CAPÍTULO 12

NOTA: la encuesta a continuación tiene por objeto insistir en que el ser humano no decide cómo es la naturaleza sino los actos que él ejecuta. Acerca de la naturaleza se puede formar una opinión determinada, pero nunca decidir como ella es. Para responder acertadamente este test debe tenerse en cuenta que el verbo «decidir» debe interpretarse en el sentido estricto al que estamos aludiendo.

Tras varios meses de juicio el juez emite veredicto condenando a Pérez a cien años de cárcel, por considerarse probado que asesinó a la viejecita. Hasta ese momento la prensa no se atrevía a llamar a Pérez «asesino», pues le era otorgada la presunción de inocencia. Pero a partir del fallo del juez los periódicos publican que ahora ya se tiene certeza de que Pérez es el asesino.

1. El veredicto implica que realmente Pérez asesinó a la viejecita.
2. El juez *decide* que Pérez asesinó a la viejecita.
3. El Juez *piensa*, cree, está convencido de que Pérez asesinó a la viejecita.
4. El juez *decide* que Pérez se vaya a la cárcel.

Veamos ahora el ejemplo ya conocido en el que estudiamos 4 presuntos anticancerígenos (AC), cada uno se los cuales se prueba en una muestra de 40 ratas de una cepa que desarrolla cáncer espontáneamente en el 60% de los casos. He aquí los resultados, el valor P del test y los intervalos de confianza.

Fármaco	Núm. de ratas con cáncer en la muestra	% de ratas con cáncer en la muestra	Valor P IC al 99%
A	5	12,5%	0,0000000003 3%-32%
B	18	45%	0,039 25%-66%
C	19	47,5%	0,074 27%-68%
D	23	57,5%	0,436 36%-77%

5. Decidimos que «A» es AC.
6. Estamos convencidos de que «A» es AC.
7. Habiendo decidido considerar estadísticamente significativos los resultados que den valor P del test menor de 0,01, concluimos que «B» no es AC.
8. Habiendo decidido considerar estadísticamente significativos los resultados que den valor P del test menor de 0,05, concluimos que «B» es AC.
9. Habiendo decidido considerar estadísticamente significativos los resultados que den valor P del test menor de 0,10, concluimos que «C» es AC.
10. No podemos pronunciarnos acerca de si «B» y «C» son o no son AC.
11. No podemos pronunciarnos acerca de si «D» es o no es AC.
12. Decidimos que «D» no es AC.

ENCUESTA DE AUTOEVALUACIÓN DEL CAPÍTULO 14

Se sabe que la media poblacional de la concentración en suero de la proteína R8 en humanos normales es 450 y la desviación estándar es 30. En las personas con trisomía 18 (una alteración cromosómica cuyos detalles no necesitamos conocer para realizar los razonamientos que siguen y que afecta al 20% de las personas) la media poblacional de R8 es 600. Tenemos tres pacientes A, B y C, a los que medimos la R8 con intención de que ello nos ayude a saber si tienen esa alteración genética. Para cada uno de ellos planteamos la H_0 que dice que no tiene la alteración, es decir, es normal, medimos R8 y calculamos el valor P del test.

$$\text{«A»}: R8 = 480 \rightarrow P_{UNILATERAL} = 0,159$$

$$\text{«B»}: R8 = 500 \rightarrow P_{UNILATERAL} = 0,047$$

$$\text{«C»}: R8 = 570 \rightarrow P_{UNILATERAL} = 0,00003$$

1. De cada 1.000 personas normales, 159 tienen R8 igual o mayor de 480.
2. De cada 1.000 personas normales, 47 tienen R8 igual o mayor de 500.
3. De cada 100.000 personas normales, 3 tienen R8 igual o mayor de 570.
4. De cada 1.000 personas normales, 159 tienen R8 igual o menor de 480.
5. De cada 1.000 personas normales, 47 tienen R8 igual o menor de 500.
6. De cada 100.000 personas normales, 3 tienen R8 igual o menor de 570.
7. De cada 1.000 personas con trisomía, 159 tienen R8 igual o mayor de 480.
8. De cada 1.000 personas con trisomía, 47 tienen R8 igual o mayor de 500.
9. De cada 100.000 personas con trisomía, 3 tienen R8 igual o mayor de 570.
10. 0,159 es la probabilidad de que «A» tenga trisomía 18.
11. 0,047 es la probabilidad de que «B» tenga trisomía 18.
12. 0,00003 es la probabilidad de que «C» tenga trisomía 18.

13. 0,159 es la probabilidad de que «A» *no* tenga trisomía 18.
14. 0,047 es la probabilidad de que «B» *no* tenga trisomía 18.
15. 0,00003 es la probabilidad de que «C» *no* tenga trisomía 18.
16. 0,159 es la probabilidad de que «A» tenga $R8 = 600$.
17. 0,047 es la probabilidad de que «B» tenga $R8 = 600$.
18. 0,00003 es la probabilidad de que «C» tenga $R8 = 600$.
19. 0,159 es la probabilidad de que «A» tenga $R8 = 450$.
20. 0,047 es la probabilidad de que «B» tenga $R8 = 450$.
21. 0,00003 es la probabilidad de que «C» tenga $R8 = 450$.
22. 0,159 es la probabilidad de que «A» tenga $R8 = 600$, si tiene trisomía.
23. 0,047 es la probabilidad de que «B» tenga $R8 = 600$, si tiene trisomía.
24. 0,00003 es la probabilidad de que «C» tenga $R8 = 600$, si tiene trisomía.
25. 0,159 es la probabilidad de que «A» tenga $R8 = 450$, si tiene trisomía.
26. 0,047 es la probabilidad de que «B» tenga $R8 = 450$, si tiene trisomía.
27. 0,00003 es la probabilidad de que «C» tenga $R8 = 450$, si tiene trisomía.
28. 0,159 es la probabilidad de que «A» tenga $R8 = 600$, si es normal.
29. 0,047 es la probabilidad de que «B» tenga $R8 = 600$, si es normal.
30. 0,00003 es la probabilidad de que «C» tenga $R8 = 600$, si es normal.
31. 0,159 es la probabilidad de que «A» tenga $R8 = 450$, si es normal.
32. 0,047 es la probabilidad de que «B» tenga $R8 = 450$, si es normal.
33. 0,00003 es la probabilidad de que «C» tenga $R8 = 450$, si es normal.

34. 0,159 es la probabilidad de que «A» tenga $R8 > 450$, si es normal.
35. 0,047 es la probabilidad de que «B» tenga $R8 > 450$, si es normal.
36. 0,00003 es la probabilidad de que «C» tenga $R8 > 450$, si es normal.
37. 0,159 es la probabilidad de que «A» tenga $R8 > 480$, si es normal.
38. 0,047 es la probabilidad de que «B» tenga $R8 > 500$, si es normal.
39. 0,00003 es la probabilidad de que «C» tenga $R8 > 570$, si es normal.
40. 0,159 es la probabilidad de que «A» tenga $R8 > 480$, si tiene trisomía.
41. 0,47 es la probabilidad de que «B» tenga $R8 > 500$, si tiene trisomía.
42. 0,00003 es la probabilidad de que «C» tenga $R8 > 570$, si tiene trisomía.

Soluciones de las encuestas de autoevaluación

A continuación se dan los números correspondientes a las afirmaciones verdaderas de cada una de las encuestas propuestas:

Encuesta previa 1: 1 – 7 – 9 – 12 – 14.

Encuesta previa 2: 2 – 5 – 6 – 9 – 12.

Encuesta previa 3: 1 – 3 – 8 – 10.

Encuesta del Capítulo 4: 1 – 2 – 5 – 9 – 12 – 15 – 18 – 19 – 21.

Encuesta del Capítulo 5: 2 – 3 – 5 – 6 – 7 – 8 – 9 – 10 – 11 – 13 – 15 – 18.

Encuesta del Capítulo 7: 3 – 5 – 9 – 16 – 22 – 25.

Encuesta del Capítulo 9: 1 – 3 – 11 – 12 – 16.

Encuesta del Capítulo 10: 1 – 3 – 4 – 5 – 6 – 7 – 9 – 12 – 13 – 15.

Encuesta del Capítulo 11: 4 – 6 – 10 – 12 – 18 – 19.

Encuesta del Capítulo 12: 3 – 4 – 6 – 10 – 11.

Encuesta del Capítulo 14: 1 – 2 – 3 – 37 – 38 – 39.

Comentarios del Profesor Rafael Romero Villafranca

NOTA: *Los comentarios que siguen fueron hechos por el Prof. Romero antes de leer la presente obra, no son un intento por su parte de resumirla después de haberla leído. La coincidencia entre sus criterios aquí expuestos y las tesis defendidas en esta obra es un fuerte aval para las mismas, no menos valioso que las citas de otros estadísticos internacionalmente reconocidos recogidas en el capítulo 2.*

«Me gustaría resaltar, desde la óptica de un estadístico industrial centrado en el control y la mejora de procesos, algunas ideas relacionadas con los temas abordados en esta obra».

1. En primer lugar, es esencial comprender bien la naturaleza de las «hipótesis nulas» H_0 que se pretenden contrastar estadísticamente. En contextos científicos la H_0 suele reflejar el estado actual de conocimiento (quizás sería más preciso decir de desconocimiento) sobre la cuestión en estudio. Por ejemplo: la eficacia de cierto fármaco frente a una determinada enfermedad no está probada o es desconocida; la posición intelectual de partida en la investigación, recogida en la H_0 , es seguir admitiendo que el fármaco no es eficaz, a no ser que haya en los datos recogidos una «evidencia fuerte» en contra de dicha hipótesis. En contextos industriales la H_0 refleja frecuentemente una actitud de prudente escepticismo. Por ejemplo: no nos creemos que un nuevo proceso B sea mejor que el proceso tradicional A, a no ser que en los datos experimentales

haya una «fuerte evidencia» contra la H_0 de que los dos procesos son iguales.

2. Y, en los ejemplos anteriores, ¿qué debe entenderse por «fuerte evidencia» contra H_0 ? Esto es precisamente lo que cuantifica el «p-value», que —bajo ciertos supuestos razonables sobre la distribución de las variables estudiadas— mide la probabilidad que existiría, en caso de ser cierta H_0 , de obtener unos datos que discreparan de dicha hipótesis tanto o más que los datos obtenidos. En este sentido, la utilización de unos valores límites o «críticos» para el p-value (como el 5% o el 1%), que separan los resultados «significativos» de los «no significativos» no es más que un anacronismo, reflejo de épocas en las que el cálculo exacto de estos p-values se hayaba fuera del alcance del investigador, que sólo podía hacerse una idea al respecto comparando los valores por él obtenidos con los reflejados en unas tablas que generalmente se limitaban a estos dos niveles. Es el p-value, por tanto, lo que refleja el grado de evidencia de unos resultados contra la H_0 y, en consecuencia, lo que debería acompañar al análisis de dichos resultados, y no sólo la constatación de si resulta superior o inferior al 5%. ¿Qué diferencia hay, en la práctica, entre un p-value del 4.9% o del 5.1%?
3. Otro error muy frecuente, es la confusión entre significación estadística e importancia práctica. El problema es de naturaleza semántica, y deriva de utilizar, para designar un concepto técnico estadístico concreto, un vocablo —«significativo»— que tiene un sentido diferente en el lenguaje habitual. Si la diferencia entre los rendimientos medios de dos procesos es «muy significativa estadísticamente», la interpretación práctica correcta es que es casi seguro que dicha diferencia no es nula, y no necesariamente que la diferencia en cuestión sea muy importante. En este sentido, el cálculo del intervalo de confianza para el efecto en cuestión es mucho más informativo que la simple constatación de si dicho intervalo contiene o no al cero, que en el fondo es lo que hace el test de hipótesis. Si el grado de incidencia de cierta patología en un colectivo es el 20%, y un estudio demuestra que la utilización preventiva de cierto fármaco lo reduce al $19.9\% \pm 0.01\%$, o sea si el intervalo, para un cierto nivel de confianza, de la reducción

en el porcentaje de incidencia es [0.09% - 0.11%], la reducción será muy significativa estadísticamente, pero posiblemente de nula importancia práctica.

4. Por otra parte, el que un determinado test estadístico no rechace una H_0 , no significa que se haya demostrado que dicha hipótesis nula es cierta, sino sólo que la misma es compatible con los datos observados, como lo serían probablemente también muchas otras hipótesis alternativas. Nuevamente el intervalo de confianza es más informativo, a efectos prácticos de ayudar a tomar una decisión, que el simple resultado del test de hipótesis. Si el rendimiento medio de un proceso es 100, y el intervalo de confianza para el incremento en la media originado por un posible cambio en estudio es $[-0,1 +0,2]$, dicho incremento no será significativo (es posible que sea cero), pero además sabemos que, como mucho, sería de 0,2, lo que posiblemente no tenga ningún interés práctico. Por el contrario, si dicho intervalo fuera $[-20 +50]$, la conclusión práctica sería diferente: es posible que el cambio no mejore nada (el incremento en la media puede que sea cero), pero también es posible que implique una mejora importante (+50) o un empeoramiento sensible (-20); el tema deberá, por tanto, estudiarse más a fondo, posiblemente mediante una experiencia más precisa.
5. En el campo de la investigación científica, el que unos resultados no lleguen a ser significativos estadísticamente (entendido ello de la forma habitual, como que el p-value sea superior al 5%) no significa necesariamente que no merezcan ser publicados, obviamente con las matizaciones pertinentes, especialmente si los efectos constatados van en el sentido que cabría esperar por las hipótesis de trabajo avanzadas en la investigación. Es posible que la no significación se deba sólo a un número insuficiente de datos, originado a veces por el elevado coste de estos estudios, pero que estos resultados, acumulados con otros obtenidos por otros equipos que trabajan sobre el tema, permitan llegar a la comunidad científica a conclusiones estadísticamente significativas sobre el tema.
6. Finalmente hay que resaltar la importancia que, para la obtención de conclusiones estadísticas correctas, tiene la constatación de

que éstas se basan en modelos estadísticos adecuados para los datos. En particular es esencial la constatación, posiblemente mediante sencillos métodos gráficos, de la ausencia de datos anómalos y de la pertinencia de dichos modelos. ¡Cuántas veces, en nuestra experiencia personal, unos resultados aparentemente muy significativos de un análisis estadístico, se debían exclusivamente a una observación anormal que había pasado desapercibida! El análisis gráfico de los «residuos» debería ser una práctica ineludible en cualquier estudio estadístico, y las revistas científicas deberían ser más exigentes al respecto, en vez de la preocupación obsesiva que algunas muestran por el mítico 5%.

7. La exposición de los fundamentos lógicos de la Inferencia Estadística en forma intuitiva no necesita un aparato matemático sofisticado, pero mi experiencia docente a lo largo de cuatro décadas me ha enseñado lo difícil que resulta para muchas personas entender la naturaleza de los razonamientos estadísticos, aunque se presenten desprovistos de formalismos matemáticos. Y es que dificultad conceptual no es sinónimo de complejidad matemática.